Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The effects of privacy-non-invasive interventions on cheating prevention and user experience in unproctored online assessments: An empirical study

Suvadeep Mukherjee[a,*], Björn Rohles[a,1], Verena Distler[a,2], Gabriele Lenzini[b], Vincent Koenig[a]

[a] Human-Computer Interaction Research Group, University of Luxembourg, L-4366, Esch-sur-Alzette, Luxembourg
[b] SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, L-4365, Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Keywords:
Unproctored cheating prevention
Remote test
Honor code
User experience (UX)
Self-efficacy

ABSTRACT

Preventing cheating without invading test-takers' privacy in high-stakes online summative assessments poses a challenge, especially when the assessment is remote and unproctored. We conducted a between-subjects experiment (N = 997) in a realistic online test simulation to investigate the effects of three privacy-non-invasive anti-cheating interventions (honor code reminder, warning message, and monitoring message) on cheating prevention from a user-centered perspective. The quantitative results indicated that, compared to a control condition, displaying a honor code reminder during an online test worked best in lowering the odds of cheating. None of the interventions affected user experience and test-taking self-efficacy significantly. Further open-ended questions revealed that interventions can cause distraction which in turn could potentially evoke negative emotions. The decision to cheat was influenced by the extent to which interventions conveyed that cheating is wrong and also by test-takers' perception of getting caught if they cheated. We derived recommendations for a fair and cheating-preventive unproctored online assessment for researchers and practitioners.

1. Introduction

Ensuring the integrity, validity, and fairness of online summative assessments is a priority not only for schools and universities (Kam et al., 2018; Ranger et al., 2020), but also for other high-stakes scenarios such as admission to higher education (Yucas, 2015), promotional exams in organizations (Mitchell et al., 2018), and teacher recruitment processes (Fontaine et al., 2020). With the advent of remote learning and teaching, cheating frauds (Yucas, 2015) have raised concerns about the credibility of online assessment[3] processes. Among the stakeholders of an online assessment, there is a discussion about the trade-off between the effectiveness of

[*] Corresponding author. Human-Computer Interaction Research Group, 11 Porte des Sciences, University of Luxembourg, L-4366, Esch-sur-Alzette, Luxembourg.

E-mail addresses: suvadeep.mukherjee@uni.lu (S. Mukherjee), info@rohles.net (B. Rohles), verena.distler@unibw.de (V. Distler), gabriele.lenzini@uni.lu (G. Lenzini), vkoenig@vo.lu (V. Koenig).

[1] Current address: Digital Learning Hub, Terres Rouges Building, 14 Porte de France, L-4360, Esch-sur-Alzette, Luxembourg.

[2] Current address: Research Group for Usable Security and Privacy, University of the Bundeswehr, Munich, Germany.

[3] The paper uses the terms assessment and test interchangeably, and both specifically pertain to the summative type.

https://doi.org/10.1016/j.compedu.2023.104925
Received 31 March 2023; Received in revised form 10 August 2023; Accepted 10 September 2023
Available online 15 September 2023

cheating prevention techniques and the perception of privacy violations among test-takers (Balash et al., 2021). Prevention techniques can be *privacy-invasive* and *privacy-non-invasive*. Privacy-invasive techniques collect test-takers' personal information by controlling their devices (Baume, 2019), whereas privacy-non-invasive techniques neither collect such information nor control devices (Pleasants et al., 2022). Privacy-invasive techniques, such as remote proctoring technology, have recently received considerable attention. Remote proctoring uses either live human invigilators or automated monitoring techniques such as eye-tracking, screen-recording, real-time biometric verification, etc. (Baume, 2019; Li et al., 2021). As these practices require taking control of the webcam, microphone, or screen, they may compromise test-takers' privacy (Balash et al., 2021; Baume, 2019), possibly increasing test anxiety (Conijn et al., 2022; Karim et al., 2014; Wuthisatian, 2020) and inhibiting test performance (Chin et al., 2017; Zeidner, 1998). Overall, from a user-centered perspective, privacy-invasive remote proctoring affects the test-taking experience (Balash et al., 2021; Karim et al., 2014) and therefore, may not be the most preferred cheating prevention solution. For the purpose of our study, we use the term user-centered perspective to encapsulate user experience and test-taking self-efficacy (details in sections 2.3.1 and 2.3.2). On the other hand, research (Corrigan-Gibbs et al., 2015; Humbert et al., 2022; Pleasants et al., 2022) has studied the impacts of interventions on cheating prevention in unproctored online assessments, which do not control the test-taking devices, and hence, do not invade test-takers' privacy. They are *honor codes, warnings about punishment, and warnings about both surveillance and punishment*. However, there are four main shortcomings in these studies.

(1) ***Selection of interventions*** - The previous studies compared different combinations of the above-mentioned privacy-non-invasive interventions. Two other potential interventions that have not been studied empirically much yet are a honor code reminder which was studied only by Bing et al. (2012), and a warning about only surveillance (i.e. monitoring) which has not yet been studied empirically to our knowledge.
(2) ***Testing conditions*** - Testing conditions across the studies lack consistency in terms of the timing of display and visual design of such interventions. For example, interventions were displayed either before a test (Bing et al., 2012; Corrigan-Gibbs et al., 2015) or both before and during a test (Pleasants et al., 2022). No studies compared interventions presented solely during a test, even though this condition could simulate the perceived fear of getting caught while cheating (Freiburger et al., 2017), similar to a proctored test. Also, the interventions did not have similar visual designs where only certain aspects were varied systematically.
(3) ***User-centered perspective*** - Anti-cheating interventions during high-stakes assessments must be designed carefully to avoid any negative impact on test-takers' performances and overall experience. The impact of privacy-non-invasive interventions has been mostly unexplored from this user experience (UX) perspective. Besides UX, test-taking self-efficacy, which is a reliable proxy for test performance (Zimmerman, 1995), is also worth investigating to predict test-takers' perceived confidence in performing well in similar test designs in the future.
(4) ***Sampling*** - To date, studies have focused on student populations from similar learning groups and educational disciplines. Although students generally are the primary target group for assessment programs, studying a population with higher demographic heterogeneity, including the students, could be a significant contribution to the literature. Non-student populations are also important for other assessment programs such as in life-long learning, recruitment processes or promotional exams in organizations.

In this study, we address these gaps and take a user-centered perspective to investigate the effectiveness of three different privacy-non-invasive anti-cheating interventions (a honor code reminder, a warning mentioning the possibility of punishment, and a message about monitoring test-takers' activities) on the cheating behavior, UX and test-taking self-efficacy. All interventions in our study have similar visual designs that are systematically varied and are presented during an unproctored online test. We performed an experiment capturing both quantitative (UX, emotional experience, and test-taking self-efficacy) and qualitative (through sentence completion method) insights to investigate the impact of the interventions.

1.1. Contribution

- We contribute empirical insights into cheating prevention from a user-centered perspective in a simulation of a high-stakes unproctored online assessment by investigating its impacts on a population with demographic heterogeneity in terms of age, gender, educational background, and past experiences with online assessments.
- We propose actionable guidelines for a cheating-preventive test design for researchers and practitioners.

2. Related work

2.1. Factors influencing cheating behavior

When it comes to high-stakes summative assessments like end-of-year university exams (Kam et al., 2018), admission tests (Yucas, 2015), promotional exams (Mitchell et al., 2018), or recruitment processes (Fontaine et al., 2020), cheating becomes a significant concern due to the importance of these evaluations (Manoharan, 2019). Research on academic dishonesty, including cheating in high-stakes tests, often explores the reasons why students or test-takers do cheat and why they do not. Much of the research (Brimble, 2016; Ghanem & Mozahem, 2019; McCabe et al., 2001) that considered the question as to why students cheat have identified a range of individual factors (such as gender, grades, self-esteem etc.) and situational factors (such as peer pressure, ethical awareness, administrative deterrence etc.) that can significantly influence a student's decision to engage in exam cheating. Amigud and Lancaster

(2019) found that academic aptitude or skills, perseverance, and self-discipline can also play a role in students' cheating decisions. On the other hand, when it comes to reasons not to cheat, Miller et al. (2011) underscored the importance of moral reasoning as a significant factor. The study highlighted the significance of internalizing integrity standards as an integral part of one's moral character. In line with this, Bandura (1990) proposed that a cheating situation must activate moral reasoning in order for it to hold moral relevance and serve as the foundation for making ethical decisions. However, fear of punishment as another deterrent to cheating may have a stronger impact than a moral appeal (Miller et al., 2011; Nagin & Pogarsky, 2003), as the moral appeal merely conveys the presence of personal monitoring but not external monitoring, suggesting a lack of external deterrent to cheating. It is also important to note that the likelihood of being caught has been found to deter students from cheating more effectively than merely stating the severity of punishment (Freiburger et al., 2017; Nagin & Pogarsky, 2003).

2.2. *Cheating prevention practices in unproctored online assessments*

Research has shown different approaches to address cheating in online assessments, both in proctored (Hu et al., 2018; Moten Jr et al., 2013) and unproctored (Bing et al., 2012; Corrigan-Gibbs et al., 2015) setups. Of late, the rising evidence of cheating in unproctored online tests has led many test-organizing institutions to turn towards remote proctoring (Conijn et al., 2022). However, this technology comes with substantial drawbacks regarding test-takers' privacy and their test experience. First, there are concerns about test-takers' privacy as proctoring companies may force them to share personal information to complete the test (Cohney et al., 2021). Second, students reported an increase in test anxiety when being proctored (Balash et al., 2021; Hylton et al., 2016), a phenomenon that can affect their performance (Chin et al., 2017; Zeidner, 1998).

Traditionally, there has been a great reliance on unproctored test administration due to increased flexibility and cost-effectiveness (Rios & Liu, 2017). A robust anti-cheating approach in unproctored tests largely involves considering test-takers' awareness of cheating, instructor vigilance, and institutional policies. Notably, high-stakes exams often employ instructional strategies like strict time constraints (Schultz et al., 2008), test questions randomization (Chen et al., 2018), no backtracking to previous questions (Ladyshewsky, 2015), and inclusion of critical-thinking tasks (Harper, 2006) to impede external help. Despite such measures, some cheaters persist in seeking outside help, prompting Dawson and Sutherland-Smith (2019) to observe that professional training in contract cheating (i.e. when test-takers outsource the assessed work) can improve instructors' ability to recognize illegitimate test responses. However, establishing instances of cheating remains complex without substantial proof, frequently dissuading instructors from reporting to the relevant authorities (Harper et al., 2019). Therefore, the shared responsibility for preventing cheating emphasizes the vital role of policy-level endeavors in nurturing an anti-cheating culture, underscored by stringent regulations and a commitment to academic integrity. Some countries have even criminalized contract cheating to prevent unauthorized external assistance (Thacker et al., 2022). Beyond these measures, other anti-cheating methods in unproctored settings address test-taker awareness by conveying anti-cheating information before or during tests, as briefly discussed below.

One of the most prevalent anti-cheating interventions in unproctored tests is the display of honor codes which shifts the responsibility of promoting and upholding academic integrity from the teacher to the students by instilling a sense of moral engagement (McCabe & Trevino, 1993). Studies (McCabe et al., 2001) showed that adopting a honor code policy in an academic institution has the potential to increase students' understanding of what cheating constitutes. Moreover, actual decreases in cheating behaviors when using honor codes have been observed (McCabe, 1993; Schwartz et al., 2013). However, Shu and Gino (2012) found that a honor code alone may not suffice; the institution's culture has to emphasize and nurture integrity. This includes reminding test-takers of their honor code to restore the moral rules that a test-taker might forget following dishonest behavior.

Warnings about specific penalties against cheating constitute another unproctored intervention. Corrigan-Gibbs et al. (2015) found that warnings against cheating are more effective than displaying a honor code. Recently, in Pleasants et al. (2022)'s study, a warning about penalties coupled with a message regarding surveillance lowered the cheating rate significantly more than displaying only a honor code. Moreover, Bing et al. (2012) found that cheating could be cut down to half if test-takers are shown a honor code reminder along with a warning about consequences, compared to displaying each separately.

Due to methodological shortcomings, the effectiveness of these interventions has not yet been conclusively demonstrated. First, studies (Bing et al., 2012; Corrigan-Gibbs et al., 2015; Pleasants et al., 2022) comparing combinations of privacy-non-invasive anti-cheating interventions are inconsistent in their timing of display. For example, Corrigan-Gibbs et al. (2015) and Bing et al. (2012) presented the interventions before a test, whereas Pleasants et al. (2022) compared a honor code presented before a test and a warning presented during a test. To simulate the perception of getting caught while cheating (Freiburger et al., 2017), as can occur in a proctored test, it is worth investigating the impact of the interventions presented only during a test. In addition, the interventions used in these studies did not have similar visual designs (e.g., variable message size, message content, etc.) that were systematically controlled. Hence a causal relationship between interventions and cheating behavior cannot be claimed. To overcome these methodological shortcomings, our study investigated privacy-non-invasive interventions presented only during a test, and with a systematically controlled similar visual design, discussed in section 4.1.2.

2.3. *User-centered parameters in online assessments*

2.3.1. *User experience*

In recent years, the impact of technology on end-users has been increasingly discussed from a user experience (UX) perspective, because it encompasses all aspects of the interaction with a digital product. According to Hassenzahl (2001), when users interact with a product, they build a subjective impression of the pragmatic and hedonic qualities of the product, which are called pragmatic and hedonic dimensions of UX, respectively. The pragmatic dimension refers to the instrumental aspects of interaction that cover usability components such as ease of use or efficiency. The hedonic dimension refers to the emotive, subjective, and temporal aspects of interaction (Hassenzahl & Tractinsky, 2006).

In the context of cheating prevention in an online assessment, the overall test-taking experience could be impacted by how the testing process is designed. For example, Baume (2019) reported concerns that remote proctoring, which may invade test-takers' privacy, could have an adverse impact on their test-taking experience. However, Butler-Henderson and Crawford (2020) noted that test-takers generally viewed online assessments positively, mainly because of the flexibility in test-taking and improved perception of test integrity offered by remote proctoring (Milone et al., 2017). Unlike the proctored test setting, studying the impacts of privacy-non-invasive interventions on test-taking experience in an unproctored test setup has mostly been unexplored. Hence, UX is important to measure in an unproctored test setup because it is a significant factor that predicts product acceptance (Mlekus et al., 2020); in our study, this is the cheating-preventive test design.

2.3.1.1. Emotional experience. Measuring usability (i.e., the pragmatic dimension of UX) is a popular practice in HCI research to assess product quality. However, prior work (Buck & Davis, 2010; Buck & Ferrer, 2012) has found that considering emotion, a hedonic dimension of UX, is also necessary for effectively influencing computer-based activities. The term 'emotional experience' is often used to signify the emotional outcomes of UX, which can be classified into positive or negative emotions. Remotely proctored exams, for instance, might increase stress (Balash et al., 2021; Karim et al., 2014) and anxiety (Hylton et al., 2016; Lilley et al., 2016; Wuthisatian, 2020) due to privacy concerns and discomfort from perceived technical problems (Sefcik et al., 2022). In the field of UX, understanding both positive and negative emotional states may prove essential to improve user interaction with products.

2.3.2. Test-taking self-efficacy

Besides user experience, a test design should facilitate test-takers in achieving their expected test performances by removing obstacles during their interaction. In an academic context, students' performances (Bandura, 1993; Honicke & Broadbent, 2016; Zimmerman, 1995) are enhanced by the belief about their own capabilities, which is termed 'self-efficacy'. Self-efficacy is a motivational construct with origins in Bandura et al. (1999)'s social-cognitive theory. It was defined as the belief that one can successfully execute the behaviors needed to produce the desired outcome. This belief influences the effort people are willing to expend and how well they cope with challenges. Hence, every intervention should be free of hindrance to that belief. Importantly, the outcomes from Bandura (1993); Credé and Phillips (2011); Honicke and Broadbent (2016); Zimmerman (1995) were based on the assumption that self-efficacy beliefs necessarily reflect a greater level of domain specificity; in these studies, the academia.

The use of a generic form of self-efficacy in a specific domain or task is a matter of debate. Self-efficacy beliefs, according to Bandura (1986) and Bandura et al. (1999), vary based on the magnitude of task difficulty, the certainty of perceived success, and the extent to which both can be generalized across situations. Most researchers have conceptualized self-efficacy as a task-specific or state-like construct (i.e., *specific self-efficacy or SSE*) by considering task difficulty and success. However, there is growing interest in *general self-efficacy (GSE)*, which refers to a trait-like dimension of self-efficacy. Eden (1988) argued that GSE has a positive impact on SSE across tasks and situations. However, other studies (Eden & Zuk, 1995; Stanley & Murphy, 1997) found that GSE failed to predict SSE. It is because the predictability of GSE and SSE depends on the scope of the performance domain (Eden, 1996). Modifications of self-efficacy indices have been developed to apply GSE to other contexts. For example, Park and Avery (2019) modified GSE indices created by Chen et al. (2001) and Schwarzer and Jerusalem (1995) and adopted them to measure behavioral aspects in the crisis management domain, while Lown (2011) created a domain-specific self-efficacy scale to measure the individual ability to deal with financial management based on Schwarzer and Jerusalem (1995, 2010)'s GSE indices.

In the literature, few studies measured test-taking self-efficacy during interventions. Some assessed test-taking self-efficacy by measuring the impact of online synchronous group interventions (Coohey & Cummings, 2019) and perceived fairness of tests (Truxillo et al., 2001). Research (Bandura, 1993; Zimmerman, 1995) has found that self-efficacy in academics is reliable for predicting test performance. In recent times, Pleasants et al. (2022) and Daffin and Jones (2018) observed that remote proctoring interventions might affect test performance, possibly as a result of increased test anxiety (Chin et al., 2017). To the best of our knowledge, there has been no previous research exploring the relationship between anti-cheating interventions and an individual's test-taking self-efficacy. Therefore, we investigate their impact on test-takers' self-efficacy in an unproctored online assessment, as this is a vital step in understanding their confidence in taking future tests.

3. Research objectives

Conducting a fair and cheating-preventive unproctored online assessment is challenging due to limited control over the test environment. An anti-cheating intervention during a high-stakes online assessment may affect the overall test-taking experience. Hence, the objective of this study is to assess the effects of three interventions on cheating prevention and to investigate any unintended negative consequences for test-takers experiences, as well as their self-efficacy related to taking the test. The research questions addressed are.

- RQ1a: What is the effect of displaying privacy-non-invasive interventions (a honor code reminder, a warning message, and a monitoring message) on cheating behavior in an unproctored online assessment?
- RQ1b: How do test-takers perceive the potential impact of interventions on their intention to cheat?

- RQ2a: What is the effect of displaying these interventions on user experience (UX) including emotional experience?
- RQ2b: What do test-takers perceive as the potential concerns by the interventions on their overall experience during test-taking?
- RQ3: What is the effect of displaying these anti-cheating interventions on test-taking self-efficacy?

4. Methodology

4.1. Research design

This study used a between-subjects experimental design to explore the relationship between four conditions (control and three privacy-non-invasive anti-cheating interventions) and six dependent variables (cheating behavior, pragmatic and hedonic UX, positive and negative emotional experience, and test-taking self-efficacy). Participants were assigned two subsequent tasks - (1) taking a standardized test in an online test portal, and (2) filling out online questionnaires (Fig. 1). The entire study took approximately 15 min per participant. Participants gave informed consent and agreed not to look up any help on the internet. In the end, they were compensated with £2. To encourage participants to try and provide the correct answers, we offered an additional performance incentive of £15 for solving all questions correctly. The entire design and all measurements were thoroughly pre-tested and refined during the pilot tests described in Appendix C.

4.1.1. Design of online test

We designed an online test interface (Fig. 2). It consisted of five unique quantitative aptitude challenges to be solved within 10 min. Questions requiring critical thinking were selected which was motivated from Harper (2006)'s study to reduce cheating. The questions were adapted to the standards of highly reputed online admission tests supervised by the Educational Testing Service (www.ets.org). Some of the recommended ways of reducing cheating were incorporated into the design, such as putting strict time limits (Schultz et al., 2008) and blocking backtracking to previous questions (Ladyshewsky, 2015). We provide the full test questions in Table B.1 in Appendix. Participants were assisted with a warning upon spending much time on a single question and were also motivated by displaying a performance-based incentive in the test interface. They also could quit the test at any point.

4.1.2. Conditions in online test

Participants were randomly assigned to one of the four groups: *honor code reminder group* (honor code was reminded), *warning group* (a warning message with probable punitive measures was displayed), *monitoring group* (a live monitoring message was displayed) and *control group* (no intervention was presented).

The displayed interventions varied experimentally in two dimensions: *directive* and *precept* (Fig. 3). The *directive* statement varied over the three treatment conditions syntactically and was kept relatively standardized semantically by asking participants not to look for any help on the internet. This was intended to minimize the subjective interpretation of cheating by the participants (Barnhardt, 2016). *Precept*, on the other hand, conveyed the type of interventions. Other design aspects (e.g., size, icon, color etc.) were held constant. The text of the intervention was kept short to facilitate test-takers comprehending it quickly.

The treatment groups saw the respective interventions only between question 2 and 5 (Fig. 1). The first question was presented without an intervention in order to motivate them towards problem-solving by avoiding possible negative affective reactions at the

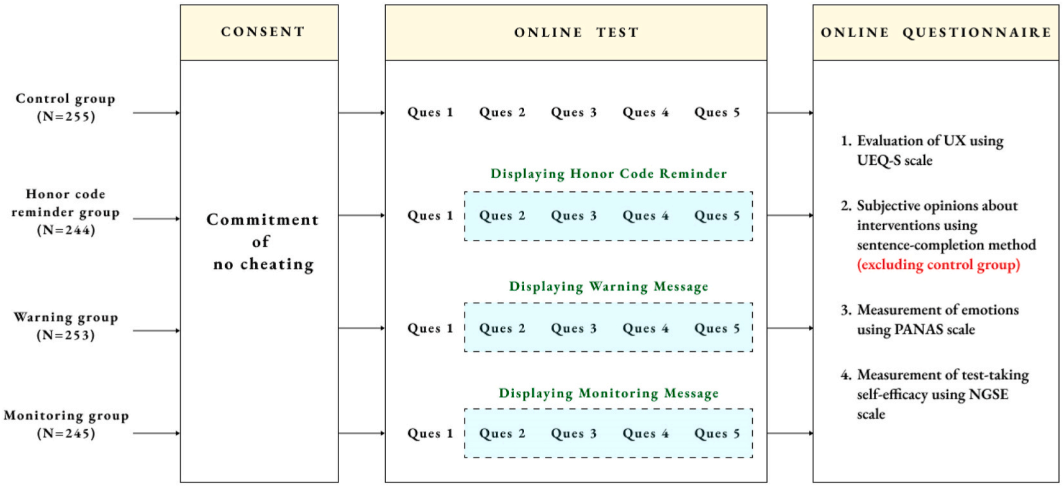
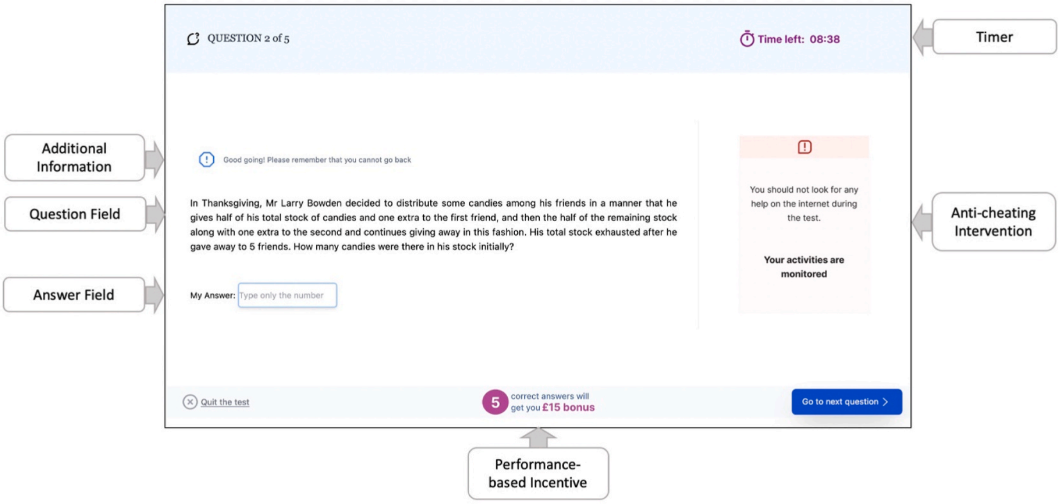

CONSENT
ONLINE TEST
ONLINE QUESTIONNAIRE
Control group (N=255)
Honor code reminder group (N=244)
Warning group (N=253)
Monitoring group (N=245)
Commitment of no cheating
Ques 1 Ques 2 Ques 3 Ques 4 Ques 5
Displaying Honor Code Reminder
Displaying Warning Message
Displaying Monitoring Message
1. Evaluation of UX using UEQ-S scale
2. Subjective opinions about interventions using sentence-completion method (excluding control group)
3. Measurement of emotions using PANAS scale
4. Measurement of test-taking self-efficacy using NGSE scale

Fig. 1. Overall design of the experiment - (1) participants committed not to look for answers on the internet in an online consent form prior to the study, (2) whoever gave their consent took an online test of five quantitative aptitude questions, (3) anti-cheating interventions are displayed in the test interface according to the group assigned randomly, (4) Finally, they filled out questionnaire on UX, open-ended questions regarding their subjective opinions about the interventions, questionnaire on emotional experience followed by test-taking self-efficacy questionnaire.



QUESTION 2 of 5

Time left: 08:38

Good going! Please remember that you cannot go back

In Thanksgiving, Mr Larry Bowden decided to distribute some candies among his friends in a manner that he gives half of his total stock of candies and one extra to the first friend, and then the half of the remaining stock along with one extra to the second and continues giving away in this fashion. His total stock exhausted after he gave away to 5 friends. How many candies were there in his stock initially?

My Answer: Type only the number

You should not look for any help on the internet during the test.

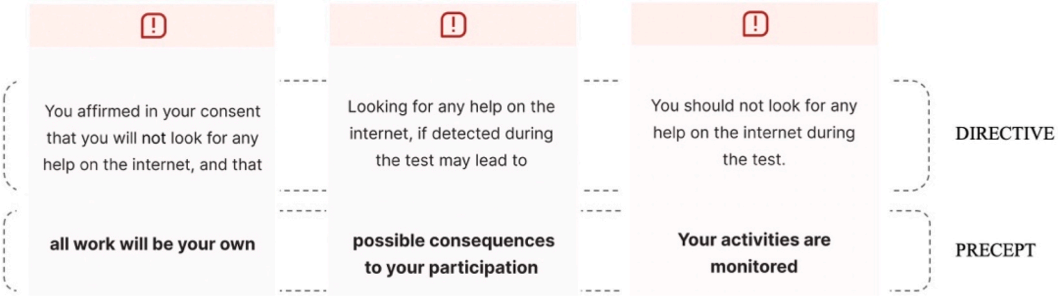
Your activities are monitored

Quit the test

5 correct answers will get you £15 bonus

Go to next question

Fig. 2. Snapshot of different components of the test interface.



You affirmed in your consent that you will not look for any help on the internet, and that

all work will be your own

Looking for any help on the internet, if detected during the test may lead to

possible consequences to your participation

You should not look for any help on the internet during the test.

Your activities are monitored

Fig. 3. Three different privacy-non-invasive anti-cheating interventions: reminding honor code (left), warning message (middle), and monitoring message (right).

beginning. From the second question, the interventions for the treatment groups appeared 10 s after the question to capture the attention of test-takers in line with the findings by Theeuwes et al. (1998).

4.1.3. *Design of online questionnaire*

After completing the test, participants were instructed to complete an online questionnaire, which is described in detail in section 4.2. In the questionnaire, a standardized user experience scale was provided for the participants to complete, with the help of a screenshot of the test system (similar to Fig. 2) to aid in recalling their overall experience. Participants who underwent the intervention were also asked to provide their subjective opinions about the interventions through a separate set of open-ended questions. Additionally, they were required to fill out a standardized scale to measure their emotions followed by a self-efficacy scale specifically designed for the test-taking context. Note that the inclusion of open-ended questions was specifically intended to gather information and infer the perceived causality regarding the impact of interventions on cheating behavior, user experience, and emotions, in particular. Participants assigned to the treatment groups were obligated to respond to the open-ended questions as illustrated in Fig. 1.

4.1.4. *Design of 'trap' websites*

Measuring test cheating accurately is challenging due to accuracy in detection and potential false accusations. Previous studies used different methods to detect cheating, such as analyzing text similarity (Humbert et al., 2022), detecting unusual incorrect answers (Schultz et al., 2022), analyzing action logs (Pleasants et al., 2022) or tracking cookies of the test-taking devices (Corrigan-Gibbs et al., 2015). However, these studies involved a certain level of subjective interpretation by researchers due to the detection methods used, which led to errors, namely false positives and false negatives. To minimize errors and increase transparency, experiments should record evidence when test-takers cheat.

Our study used an experimental design similar to Corrigan-Gibbs et al. (2015), who measured cheating rates by posting test questions on a public website without answers and providing cheaters with web buttons to click for answers. However, they couldn't detect cheaters who used a second device to search for answers. To address this problem and generally improve cheating measurement certainty, we provided answers through web buttons on three public 'trap' websites indexed by search engines. We distributed all test questions on these websites to reduce suspicion among test-takers. The answers were generated uniquely per visitor and had a range of realistic values (Table B.1 in Appendix). Section 4.2.1 outlines the procedure for detecting cheating through the utilization of trap websites, which we discuss in detail.

4.2. *Measurements*

4.2.1. *Measuring cheating behavior*

During the test, participants' responses were recorded in the database along with pseudonymized IP addresses, question numbers, device types, browser types and local system times. If a participant retrieved answers from a 'trap' website, the corresponding entries were also recorded (Fig. 4), allowing reliable identification of cheating behavior even if test-takers switched to another device to search for answers. The flowchart in Fig. 5 explains the procedure. The procedure involved checking for answer similarity between the website-generated ones and test-takers' entries in the test interface, similarity in pseudonymized IP addresses (device accessing test and device accessing 'trap' websites) and local system times[4] matching, followed by a manual validation. For question 5, having a smaller range of website-generated answers, we separated each iteration of generated values by the 'trap' websites and matched it with the entered answers that were recorded within the corresponding timeframe. The decision about cheating then followed the procedure stated in Fig. 5. Cheating behavior was measured by determining if a participant cheated at least once between question 2 and 5. A later section (5.5) examines participants' strategies, including changes in browsers or devices used for cheating.

4.2.2. *Measuring user experience (UX)*

AttrakDiff (Hassenzahl et al., 2003), UEQ (Laugwitz et al., 2008), and meCUE (Minge & Riedel, 2013) are the three most recognized standardized questionnaires for UX evaluation. While AttrakDiff and UEQ offer a subjective evaluation of pragmatic and hedonic qualities, UEQ is more balanced (Laugwitz et al., 2008). MeCUE provides self-reported assessments of product perceptions, emotions, consequences, and overall evaluations. As the study aimed to measure emotions separately with more dimensions than meCUE, UEQ was used as the UX evaluation measure.

We evaluated UX with the standardized 8-item short version of the UEQ or UEQ-S (Schrepp et al., 2017) because it takes less time to fill and provides sufficient psychometric quality for our purposes, compared to the full version (Schrepp et al., 2017). Each dimension of UEQ-S (pragmatic and hedonic) is measured with a 7-point semantic differential scale with four items. We generated two mean values for UX Pragmatic and UX Hedonic.

4.2.3. *Measuring emotional experience*

We measured the emotional aspect of user experiences specifically as its hedonic dimension. Traditional evaluation methods rely on self-reported data to measure emotional states. The Positive and Negative Affect Schedule (PANAS) scale (Watson et al., 1988) is a widely used instrument for this purpose. Positive and negative affect are two independent unipolar dimensions that encompass all affective states with positive (active, proud, enthusiastic, etc.) and negative (upset, afraid, nervous, etc.) valences, respectively.

We measured the emotional component of experiences with the standardized 20-item PANAS scale. Each dimension (positive and negative) is measured with a 5-point Likert scale *(not at all* to *extremely*) with ten items each. We computed mean scale values for both positive and negative emotions.

4.2.4. *Measuring test-taking self-efficacy*

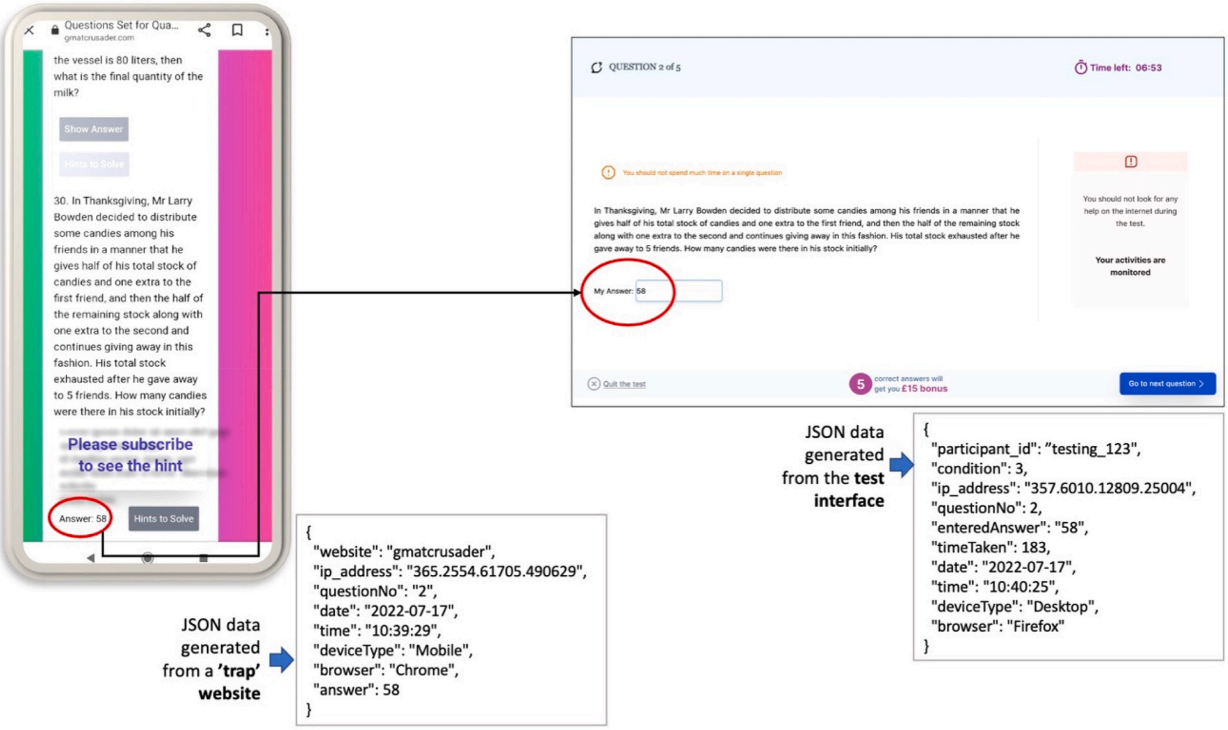
The most widely used general self-efficacy (GSE) indices are developed by Sherer et al. (1982) and Schwarzer and Jerusalem (1995). Later, Chen et al. (2001) developed a modified self-efficacy index based on the GSE indices, titled the new general self-efficacy (NGSE) index, and argued that the reliability and validity of the NGSE index are higher than that of other GSE indices.

We used an ad-hoc measure of test-taking self-efficacy based on the standardized 8-item NGSE (Chen et al., 2001). The items were modified to the study context (Table B.3 in Appendix) to make it a domain-specific measure as discussed in Section 2.3.2. Participants rated how much they *agreed* or *disagreed* with each item on a 5-point Likert scale. We computed mean scale values by averaging the respective items. The Internal consistency reliability for the scale to measure test-taking self-efficacy is calculated as 0.95 using Cronbach's alpha.

4.2.5. *Qualitative questions*

In addition to the above, we used sentence-completion method (Kujala et al., 2014) to gather qualitative insights on participants' views about the interventions. This technique yields a broader range of responses and helps understand users' thoughts, feelings, experiences, and motives (Doherty & Nelson, 2010; Donoghue, 2000). Participants in the treatment groups completed five sentences related to the intervention messages, including questions about how the interventions affected their behavior and emotions. We also measured their perceptions of the interventions' impact on others (Table B.2 in Appendix).

[4] The time difference should be positive, as cheaters enter answers later than they fetch from the 'trap' websites. We assume that the cheaters should not take more than 2 min (i.e. the average time to answer each question) to enter a fetched answer.



Questions Set for Qua...
gmatcrusader.com

the vessel is 80 liters, then what is the final quantity of the milk?

Show Answer

Hints to Solve

30. In Thanksgiving, Mr Larry Bowden decided to distribute some candies among his friends in a manner that he gives half of his total stock of candies and one extra to the first friend, and then the half of the remaining stock along with one extra to the second and continues giving away in this fashion. His total stock exhausted after he gave away to 5 friends. How many candies were there in his stock initially?

Please subscribe to see the hint

Answer: 58

Hints to Solve

QUESTION 2 of 5

Time left: 06:53

You should not spend much time on a single question

In Thanksgiving, Mr Larry Bowden decided to distribute some candies among his friends in a manner that he gives half of his total stock of candies and one extra to the first friend, and then the half of the remaining stock along with one extra to the second and continues giving away in this fashion. His total stock exhausted after he gave away to 5 friends. How many candies were there in his stock initially?

My Answer: 58

You should not look for any help on the internet during the test.

Your activities are monitored

Quit the test

5 correct answers will get you £15 bonus

Go to next question

```
{
 "participant_id": ”testing_123",
 "condition": 3,
 "ip_address": "357.6010.12809.25004",
 "questionNo": 2,
 "enteredAnswer": "58",
 "timeTaken": 183,
 "date": "2022-07-17",
 "time": "10:40:25",
 "deviceType": "Desktop",
 "browser": "Firefox"
}
```

```
{
 "website": "gmatcrusader",
 "ip_address": "365.2554.61705.490629",
 "questionNo": "2",
 "date": "2022-07-17",
 "time": "10:39:29",
 "deviceType": "Mobile",
 "browser": "Chrome",
 "answer": 58
}
```

Fig. 4. (Left) A uniquely generated answer in a ‘trap’ website (shown here on a mobile) is entered in the (right) test system (shown here on a desktop) and the overview of JSON data for both processes in MongoDB cloud; the IP addresses are strictly pseudonymized.

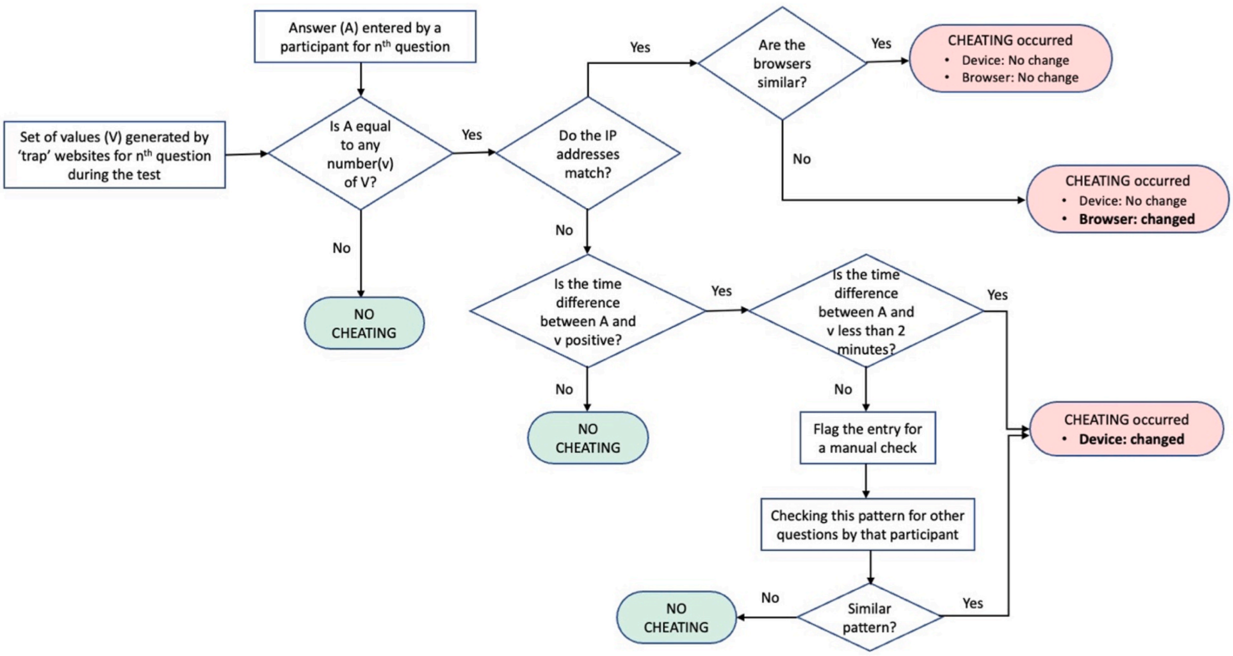

Answer (A) entered by a participant for nth question
Set of values (V) generated by ‘trap’ websites for nth question during the test
Is A equal to any number(v) of V?
Yes
No
NO CHEATING
Do the IP addresses match?
Yes
Are the browsers similar?
Yes
CHEATING occurred
• Device: No change
• Browser: No change
No
CHEATING occurred
• Device: No change
• **Browser: changed**
No
Is the time difference between A and v positive?
No
NO CHEATING
Yes
Is the time difference between A and v less than 2 minutes?
Yes
CHEATING occurred
• **Device: changed**
No
Flag the entry for a manual check
Checking this pattern for other questions by that participant
Similar pattern?
No
NO CHEATING
Yes

Fig. 5. The flowchart represents the processes followed in decision-making about the incidence of cheating during test-taking.

4.3. Recruitment and participants

We recruited 997 UK-based participants (excluding pilot study participants) from Prolific, a reliable crowd-working platform providing higher quality data (Peer et al., 2022) than other popular platforms such as MTurk. Participants were randomly assigned to one of four conditions (control group: 255, honor code reminder group: 244, warning group: 253, and monitoring group: 245). The data collection took one day in July 2022. The sample was non-representative, with 49.3% female, 49.7% male, and 0.9% non-binary participants. On average, participants were 38 years old (SD = 13) with diverse educational backgrounds, and around 70% had a university degree. Approximately 18% had taken more than 20 online tests in the last two years, while 20% had not taken any.

4.4. Ethical consideration

Participants were given a digital informed consent and a study information sheet before the study. To avoid bias regarding measuring cheating behavior, we informed them that the study aimed to improve user interaction with the test interface and would reveal the study objectives during the debriefing. We followed GDPR practices and informed participants about data collection, storage, and opt-out opportunities. After the study, we debriefed participants on the detailed purpose and approach for measuring cheating behavior. IP addresses were strictly pseudonymized, and no disciplinary action was taken against cheating participants. Every participant was compensated fairly, and the online test included questions that were moderately difficult with the promise of a bonus. However, none of the participants were able to answer all the questions correctly, resulting in them not receiving the bonus. Note that the cheaters could never get all questions correct as the websites never provided the correct answer. Considering the delayed debriefing, we pre-screened only those participants who agreed to be involved in a deception study. The university's ethics committee reviewed and approved our experimental protocol.

4.5. Data analysis

4.5.1. Quantitative analysis

To account for the binary nature of *cheating behavior* (if participants cheated or not), we used a binary logistic regression model. For other dependent variables (*UX Pragmatic, UX Hedonic, positive emotion, negative emotion, test-taking self-efficacy),* we conducted Ordinary Least Squares (OLS) regressions. The regressions were conducted using robust standard errors to account for heteroskedasticity. We estimated three separate models for our dependent variables. At first, we estimated the overall effect of anti-cheating interventions on each of our dependent variables in a model. We then estimated the individual effect of different anti-cheating interventions on each of our dependent variables in a second model. If we found significance in model 2, we compared the effect of different interventions on the dependent variables between each other in the third model. For models 1 and 2, we used the control condition as a baseline for regressions. Statistical analyses have been conducted using STATA v17. For binary logistic regression, the odds ratios[5] are reported in place of regression coefficients to interpret the effect size. In Equation (2), |OR-1| informs how much less or more the impact will be under the influence of treatment in comparison to that of control. For OLS regression, effect sizes are reported following Cohen's convention.

$$odds\ ratio\ (OR) = \frac{odds(cheating)_{treatment}}{odds(cheating)_{control}} = \frac{\frac{prob.(cheating)_{treatment}}{prob.(no\ cheating)_{treatment}}}{\frac{prob.(cheating)_{control}}{prob.(no\ cheating)_{control}}} \tag{1}$$

$$odds(cheating)_{treatment} = odds(cheating)_{control} + (OR - 1) * odds(cheating)_{control} \tag{2}$$

We checked for bivariate correlations (Table A.1 in Appendix) between the four conditions and respondent characteristics (age, gender, university degree, and past experiences with online tests) before running regressions. The correlations were close to zero and not significant, except for age and university degree, which had small correlations with some conditions. However, these correlations were also close to zero, indicating successful randomization.

4.5.2. Qualitative analysis

For qualitative analysis of the responses regarding participants' perception about the intention of the interventions and the perceived impacts on their behavior and emotions, we used an inductive coding process in MAXQDA (ver. 2022). Two coders (first and second author) individually analyzed the data and developed an open coding scheme (Creswell & Poth, 2016). Non-agreement cases were discussed, and codes adapted. A selective coding process was conducted by the coders together, which identifies categories (e.g., impact on UX etc.) so that the codes (e.g., distraction to test-taking) can be assigned to them. Any disagreement on assigning codes was discussed among the coders and resolved accordingly. To avoid misunderstandings or guessing impacting the results, we conducted a quality check to understand how accurately participants could remember the intervention message. Only qualitative inputs that were either correct or partially correct interpretations of the message were considered. Note that we did not measure inter-rater reliability because the data were short and straightforward (McDonald et al., 2019).

[5] 'Odds' are defined as the ratio of the probability of an outcome occurring to the probability of that outcome not occurring (Equation (1)); an Odd Ratio is the ratio of odds of an outcome between two groups.

5. Results

We will first describe the summaries of our dependent variables (*cheating behavior* in 5.1.1, *UX* and *emotional experience* in 5.2.1, and *test-taking self-efficacy* in 5.3.1), followed by statistical analyses to answer the research questions (*RQ1a* in 5.1.2, *RQ2a* in 5.2.2, and *RQ3* in 5.3.2) stated in section 3. We also include a section (5.4) stating the qualitative analysis specifically for dependent variables *cheating behavior* and *user experience* answering *RQ1b* and *RQ2b* respectively. Finally, we present exploratory findings about the observed cheating strategies (5.5).

5.1. Effect of anti-cheating interventions on cheating behavior

5.1.1. Descriptive statistics of cheating behavior

We observed that participants cheated more in the control condition (21.2%) than in treatment conditions (13.5%). Within treatment groups, cheating rates were 11.9% with a honor code reminder, 13.1% with a warning message, and 15.5% with a monitoring message (Table 1).

5.1.2. Statistical analyses of effect of interventions on cheating behavior

Table 2 displays the results of logistic regression on cheating behavior with different anti-cheating interventions. Overall, the interventions had a significant negative[6] effect ($p = 0.004$), resulting in a 42% decrease in cheating odds[7] compared to control. Honor code reminders ($p = 0.006$) and warning messages ($p = 0.016$) had significant negative effects on cheating (50% and 44% decrease in odds, respectively). However, the monitoring condition did not have a statistically significant ($p = 0.103$) effect on cheating, although there was a 32% decrease in the odds of cheating from control. Wald tests showed no significant differences between the interventions.

5.2. Effect of anti-cheating interventions on user experience (UX)

5.2.1. Descriptive statistics of UX-pragmatic, UX-hedonic and emotional experience

The results presented in Table 3 indicate that, in general, the pragmatic qualities of user experience (UX) received higher ratings compared to the hedonic qualities. Furthermore, among the different treatments, the honor code reminder received higher ratings than both the warning and monitoring conditions. Ratings ranged from −0.8 to +0.8, indicating a neutral UX evaluation (Schrepp, 2019) in both control and treatment groups. In addition to UX qualities, Table 4 reveals that all interventions evoked positive emotions more frequently than negative emotions.

5.2.2. Statistical analyses of effect on UX-pragmatic, UX-hedonic and emotional experience

Table 5 presents OLS regression results on UX with different anti-cheating interventions. The overall impact of the interventions was small and statistically non-significant on both pragmatic ($p = 0.373$) and hedonic qualities ($p = 0.743$). Likewise, no individual intervention had a significant effect on either of these qualities.

Table 6 displays OLS regression results on emotional experience with different anti-cheating interventions. Similar to the findings for UX, the overall impact of the interventions on positive emotions ($p = 0.559$) and negative emotions ($p = 0.283$) was small and not statistically significant. Additionally, no individual intervention had a significant effect on either positive or negative emotions.

5.3. Effect of anti-cheating interventions on test-taking self-efficacy

5.3.1. Descriptive statistics of test-taking self-efficacy

Overall, participants had a neutral evaluation of test-taking self-efficacy (Table 7). Moreover, self-efficacy under the influence of honor code reminders was rated slightly higher than the other interventions, while the warning was rated the lowest.

5.3.2. Statistical analyses of effect of interventions on test-taking self-efficacy

Table 8 displays OLS regression results on test-taking self-efficacy with different anti-cheating interventions. Both the overall effect ($p = 0.136$) and the individual effects of the interventions were found to be statistically non-significant.

5.4. Results of qualitative analysis

Using qualitative answers, we compared the perceived impact of different anti-cheating interventions on the test-takers' cheating behavior and test-taking experiences including emotions (Table 9). These inputs provide further insights into the interventions' effects on overall user experience. We will discuss the general insights, followed by specific contrasting insights for each intervention in the subsections below.

[6] An odds ratio less than 1 signifies a negative association and more than 1 means a positive association.

[7] The effect of treatment is measured using Equation (2).

Table 1
Percentage of cheaters across conditions.

Conditions	Mean	Std. Dev.	Std. Error	Sample size
Control (no intervention)	21.2%	40.9%	2.6%	255
Honor code reminder	11.9%	32.4%	2.1%	244
Warning	13.1%	33.8%	2.1%	253
Monitoring	15.5%	36.3%	2.3%	245

Table 2
Effects of honor code reminder, warning, and monitoring interventions on cheating behavior.

	Conditions	Odds ratios with robust std. Error	95% CI
Model 1	Treatment	0.579** (0.108)	[0.401, 0.836]
	Control	1.000	
Model 2	Honor code reminder	0.502** (0.125)	[0.307, 0.820]
	Warning	0.558* (0.134)	[0.347, 0.896]
	Monitoring	0.683^{+} (0.159)	[0.432, 1.080]
	Control	1.000	

- Odds ratios of control as baseline is shown as 1 so one can better see which odds ratios of treatments are significantly different from 1
- $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Comparing effects between interventions	Results
Model 3	Honor code reminder and warning	χ^2 (1, N = 497) = 0.15, p = 0.7
	Honor code reminder and monitoring	χ^2 (1, N = 489) = 1.35, p = 0.25
	Warning and monitoring	χ^2 (1, N = 498) = 0.62, p = 0.43

Table 3
Results of standardized 8-item UEQ-S with 7-point Likert scale. The self-reported values are transformed into a −3 to +3 scale. The +3 represents the most positive experience and the −3 the most negative experience.

Conditions	UX-Pragmatic				UX-Hedonic			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Control (no intervention)	0.73	1.49	−2.75	3.00	0.39	1.22	−2.50	3.00
Honor code reminder	0.78	1.53	−3.00	3.00	0.43	1.29	−3.00	3.00
Warning	0.53	1.57	−3.00	3.00	0.32	1.31	−3.00	3.00
Monitoring	0.61	1.59	−3.00	3.00	0.33	1.27	−3.00	3.00

Table 4
Results of 20-item PANAS with 5-point Likert scale. The self-reported values are grouped into positive and negative emotion values. 1 represents very slightly or not at all evoked emotion and 5 represents extremely evoked emotion.

Conditions	Positive emotion				Negative emotion			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Control (no intervention)	2.94	0.81	1.00	4.80	1.75	0.67	1.00	4.40
Honor code reminder	2.95	0.75	1.00	5.00	1.78	0.66	1.00	4.60
Warning	2.82	0.81	1.10	5.00	1.82	0.69	1.00	4.40
Monitoring	2.96	0.79	1.00	4.80	1.79	0.69	1.00	4.40

5.4.1. Effect of anti-cheating interventions on cheating behavior

Participants had varying perceptions of the interventions' impact on their behavior. Some believed the interventions prevented them from cheating, while others did not perceive any impact. Some participants had pre-determined not to cheat, while others did not cheat because of compliance with the instructions (*P783: "I was not going to cheat as it said not to use the internet at the start, I see no point in going against instructions"*); lack of time (*P927: "I didn't want to cheat and didn't have time to do so anyway"*) and other undisclosed reasons (*P108: "I did not cheat, even though I was tempted to look up how to solve one of the problems"*).

Overall, one-third of honor code reminder group participants were pre-determined not to cheat, compared to one-fourth in the warning group and one-fifth in the monitoring group. Few participants only in the honor code reminder group understood the interventions' appeal to comply with rules (*P778: "reminding me of the importance of being honest during exams"*). Table 9 indicates that monitoring had the highest percentage (32%) of participants who admitted that the intervention deterred them from cheating (*P960: "Stopped me from being tempted to google to answers to the questions"*), whereas only 16% and 24% admitted in honor code reminders and

Table 5
Effects of honor code reminder, warning, and monitoring interventions on user experience (UX).

	Conditions	UX-Pragmatic		UX-Hedonic	
		Coefficients with robust std. error	95% CI	Coefficients with robust std. error	95% CI
Model 1	Treatment	−0.098 (0.109)	[-0.312, 0.117]	0.029 (0.089)	[-0.206, 0.147]
	Control	0.000		0.000	
Model 2	Honor code reminder	0.041 (0.135)	[-0.224, 0.306]	0.039 (0.113)	[-0.181, 0.261]
	Warning	−0.201 (0.135)	[-0.467, 0.064]	−0.068 (0.112)	[-0.288, 0.15]
	Monitoring	−0.128 (0.138)	[-0.399, 0.143]	−0.058 (0.111)	[-0.277, 0.159]
	Control	0.000		0.000	

- Coefficients of control as baseline is shown as 0 so one can better see which coeffs of treatments are significantly different from 0.
- Mean value index of UX-Pragmatic and UX-Hedonic based on 4 items each on a scale of −3 to +3.
- $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 6
Effects of honor code reminder, warning, and monitoring interventions on emotional experience.

	Conditions	Positive emotion		Negative emotion	
		Coefficients with robust std. error	95% CI	Coefficients with robust std. error	95% CI
Model 1	Treatment	−0.034 (0.058)	[-0.148, 0.081]	0.053 (0.049)	[-0.043, 0.149]
	Control	0.000		0.000	
Model 2	Honor code reminder	0.006 (0.069)	[-0.130, 0.143]	0.032 (0.059)	[-0.084, 0.151]
	Warning	-0.124^{+} (0.071)	[-0.264, 0.016]	0.071 (0.061)	[-0.047, 0.189]
	Monitoring	0.018 (0.071)	[-0.121, 0.159]	0.053 (0.060)	[-0.066, 0.173]
	Control	0.000		0.000	

- Coefficients of control as baseline is shown as 0 so one can better see which coeffs of treatments are significantly different from 0.
- Mean value index of positive and negative emotions based on 10 items each on a scale of 1 to 5.
- $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 7
Results of 8-item Test-taking Self-Efficacy statements with a 5-point Likert scale. 1 represents strongly disagreed with the statement and 5 represents strongly agreed with the statement.

Conditions	Mean	SD	Min	Max
Control (no intervention)	3.03	0.91	1.00	5.00
Honor code reminder	3.01	0.86	1.00	5.00
Warning	2.87	0.95	1.00	5.00
Monitoring	2.92	0.95	1.00	5.00

Table 8
Effects of honor code reminder, warning, and monitoring interventions on test-taking self-efficacy.

	Conditions	Coefficients with robust std. error	95% CI
Model 1	Treatment	−0.098 (0.066)	[-0.229, 0.031]
	Control	0.000	
Model 2	Honor code reminder	−0.027 (0.079)	[-0.182, 0.128]
	Warning	-0.157^{+} (0.082)	[-0.319, 0.004]
	Monitoring	−0.109 (0.083)	[-0.273, 0.054]
	Control	0.000	

- Coefficients of control as baseline is shown as 0 so one can better see which coeffs of treatments are significantly different from 0.
- Mean value index of self-efficacy based on 8 items each on a scale of 1 to 5.
- $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

warning groups respectively. Participants in the monitoring and warning groups felt more watched (27% and 24%, respectively) than those in the honor code reminder group (10%). These findings will be discussed further in section 6.2.

Participants raised concerns about intervention effectiveness. In the monitoring condition, some believed interventions prompted cheating (*P598: "… perhaps be reminded that the answers may be on the internet and therefore to search using a separate device"*). Not mentioning consequences in warning messages could also incentivize cheating (*P779: "Seems easy to bypass those checks, but I guess you have to know they are there"*). In addition, the absence of real-time tracking may not yield long-term results (*P796: "… I guess some might*

Table 9
Qualitative comparison of insights along with number of mentions across treatment conditions.

		Honor code reminder	Warning	Monitoring
Quality check	Total sample size	244	253	245
	Did not recall the intervention message correctly	59 (24%)	54 (21%)	49 (20%)
	Sample excluding who did not recall the message	185	199	196
Correctness in their comprehension	Correct comprehension of the recalled intervention message	50 (27%)	44 (22%)	90 (46%)
	Partially correct comprehension of the recalled intervention message	135 (73%)	155 (78%)	106 (54%)
Impact on behavior (subjective views)	Participants expressed their pre-determination not to cheat	63 (34%)	50 (25%)	39 (20%)
	Participants expressed that the message prevented them from cheating	29 (16%)	48 (24%)	63 (32%)
	Participants did not feel any impact	46 (25%)	53 (27%)	57 (29%)
	Participants felt distracted	43 (23%)	30 (15%)	18 (9%)
	Participants felt being watched	19 (10%)	47 (24%)	52 (27%)
Impact on emotion (subjective views)	Participants expressed positive emotions	45 (24%)	23 (12%)	46 (23%)
	Participants expressed negative emotions	113 (61%)	131 (66%)	129 (66%)
	Participants did not feel any emotional change	23 (13%)	27 (14%)	26 (13%)

ignore it if they felt it wasn't actually genuine").

5.4.2. Effect of anti-cheating interventions on emotions

Participants reported varying levels of positive emotions (motivated, convinced, proud, safe etc.) and negative emotions (worried, stressed, annoyed, scared, anxious etc.) in different interventions. About 27% of participants said they felt no impact (*P786: "it didn't make me feel anything*"), while 3% reported extreme emotional reactions (*P4: "... Paranoid that the system had registered me as cheating when I hadn't*"). Positive emotions included comfort and safety, such as "*happy that they won't let people cheat" (P35), "... I didn't feel it was aimed at me" (P163), "safe that other people can't cheat" (P656)*, while negative emotions included worry (*P102: "worried that I might be falsely accused of cheating*"), stress and annoyance (*P206: "stressed and under pressure ie. if I took too long it would think I was googling answers"; P9: "it made me feel even more annoyed*") and distrust (*P632: "like the researchers did not trust me*").

Fear was more prominent in warning and monitoring interventions as indicated by the participants, potentially related to worries about false accusations (*P689: "I was worried it had malfunctioned or incorrectly perceived my actions as an attempt to cheat*"). The likelihood of stress, distrust, and annoyance were reported across all interventions. Positive emotions were less frequently reported in the warning condition (12%) compared to the honor code reminder (24%) or monitoring condition (23%).

5.4.3. Effect of anti-cheating interventions on user experience

The qualitative sentence-completion responses revealed that around 16% of participants were distracted by the interventions (*P651: "It slowed me down because I had to spend time reading it*"). The 10-s delay may have reinforced the impression of being distracted (*P592: "... reddish window with the alert to work alone appeared a bit out of nowhere and broke my concentration*") and watched (*P129 under warning condition: "I checked to see if you were watching me on my webcam*", and *P154 under honor code condition: "like I was being*

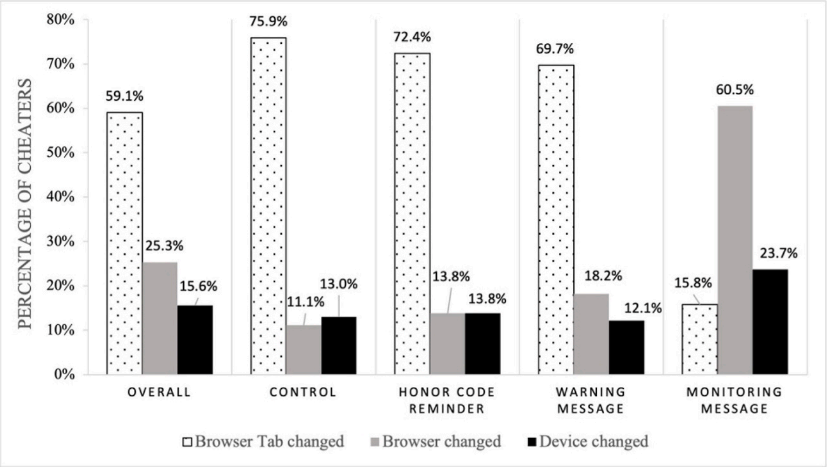


Fig. 6. Distribution of the cheating strategies adopted by cheaters overall and in different conditions.

watched"). However, some (around 9%) participants expressed positive experiences with the interventions, such as *"it assured me that the platform is fair for all takers" (P921)* and *"... it was supportive" (P792)*.

Honor code reminder participants (23%) reported being more frequently distracted than warning (15%) or monitoring (9%) participants due to message redundancy. No new information was presented, because *"it was information that had been conveyed before" (P437)* as a part of the consent process.

5.5. *Post-hoc analysis of cheating behavior*

The study not only examines the impact of different privacy-non-invasive interventions on cheating behavior but also explores the cheating strategies adopted by test-takers, specifically device and browser switching. The study aimed to address Corrigan-Gibbs et al. (2015)'s argument about cheaters switching devices to avoid detection, especially in the context of BYOD test setups. We found that 25% of cheaters changed browsers and 16% changed devices when accessing cheating websites (Fig. 6). Cheaters in the monitoring condition switched browsers significantly more often ($p = 0.000$) than the control group (Table 10), but there were no significant differences in device or browser switching in the other treatment conditions. We will return to this in the discussion section.

6. Discussion

6.1. *Rationale behind the study design*

Our research aimed to investigate the tension between the effectiveness of cheating prevention and the potential unintended harm that any anti-cheating intervention may cause to both the test-taking experience and test-taking self-efficacy. Based on literature, privacy-invasive interventions, such as remote proctoring, seek to control test-takers' devices and collect their information which may affect the user-centered parameters (i.e. user experience and test-taking self-efficacy). Hence, we opted to assess interventions that do not invade on users' privacy. Although our study design collected data to detect any deliberate attempts to cheat by changing browsers or devices during the experiment, it did not entail regulating the test-takers' devices, as is the case with remote proctoring.

In contrast to earlier investigations (Corrigan-Gibbs et al., 2015; Humbert et al., 2022; Malesky et al., 2022; Pleasants et al., 2022), our research was carried out with a substantial and varied pool of participants, having different backgrounds and levels of experience with online testing. This was important for ensuring robust statistical inferences. To achieve this, we included a sample question in the screening process, and only those who demonstrated confidence in solving critical thinking-based questions on the spot were included in the study. Our study design was successful in replicating a realistic test environment, as we implemented a "no cheating" pledge at the beginning of the test and increased the stakes by introducing a bonus.

6.2. *Interpreting the impacts of anti-cheating interventions on cheating behavior*

Our study collected empirical evidence and qualitative inputs from participants on interventions to prevent cheating during a test. The control group had a cheating rate (21.2%) that lies within the range of findings from other recent experiment-based studies (14.0% in Humbert et al. (2022), 18.7% in Malesky et al. (2022), 34.4% in Corrigan-Gibbs et al. (2015) and 50.0% in Bing et al. (2012)). Around 11.9% of participants in our study who were reminded of the honor code engaged in cheating, while 13.0% of those who received a warning and 15.5% of those who received monitoring messages were found to cheat. This contradicts other studies that showed warnings of punishment (Bing et al., 2012) or surveillance (Pleasants et al., 2022) to be more effective in preventing cheating than simply displaying a honor code. The difference in our study's results may be attributed to the fact that we reminded test-takers of the honor code during test-taking, unlike in the previous studies. We will now discuss the effects of the interventions individually.

Table 10
Effects of honor code reminder, warning, and monitoring interventions on cheating strategies adopted by cheaters.

	Conditions	Odds ratios with robust std. error	95% CI
Model 1 (Any change in either browser or device)	Treatment	3.362** (1.294)	[1.582, 7.149]
	Control	1.000	
Model 2 (Change in browser only)	Honor code reminder	1.28 (0.887)	[0.328, 4.981]
	Warning	1.778 (1.115)	[0.519, 6.081]
	Monitoring	12.267*** (6.714)	[4.195, 35.862]
	Control	1.000	
Model 3 (Change in device only)	Honor code reminder	1.074 (0.726)	[0.285, 4.042]
	Warning	0.926 (0.622)	[0.248, 3.456]
	Monitoring	2.083 (1.163)	[0.697, 6.224]
	Control	1.000	

- Odds ratios of control as baseline is shown as 1 so one can better see which odds ratios of treatments are significantly different from 1.
- $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

6.2.1. Honor code reminders on cheating behavior

Studies suggest that people desire to view themselves as ethical (Lerner, 1977), and being dishonest can cause discomfort (Klass, 1978; Shaffer, 1975), leading to a change in their attitude (Fazio & Cooper, 1983) and behavior (Baumeister & Heatherton, 1996). Our analysis showed that about one-fourth of participants recognized the honor code reminder as part of their consent. This reminder was inspired by Shu and Gino (2012)'s study on restoring moral rules through an 'appeal'. Interestingly, there were more frequent mentions of "pre-determination against cheating'' in the honor code reminder group (see Table 9). Though few participants mentioned the appeal, it may have strengthened their ethical behavior habits (McCabe & Trevino, 1993) and reduced cheating. About one-tenth of participants in this group felt watched, likely due to the delayed appearance of the message, which may have also reduced cheating.

6.2.2. Warning messages on cheating behavior

Brinthaupt (2004) suggested that an expectation-lowering procedure like warnings could reduce academic cheating by developing students' expectations about getting caught and the consequences. Miller et al. (2011) supported this idea, stating that punishment affects cheating prevention when the salience of punishment and the perception of being caught are high. In our study, more (around one-fourth) participants in this group felt like being watched than the ones who recalled the involvement of punitive measures (around one-fifth), most likely due to the delayed appearance of the message. However, unlike the honor code reminders, the punitive measures may not have activated the test-takers' moral reasoning well. This is evident from the increase in switching browsers or devices for cheating (Fig. 6) in the warning group compared to the honor code reminder group.

6.2.3. Monitoring messages on cheating behavior

Unlike other interventions, monitoring messages explicitly indicate the presence of a tracker without appealing for moral rules or including punishment. However, the presence of a monitoring message was not effective enough in reducing cheating. Cheaters in this group switched browsers (61% of the time) and devices (24% of the time) more often to hide their behavior (Fig. 6). Not mentioning the consequences of getting caught could indirectly incentivize cheating. Note that inaccurate detection of browser or device switches gives an inaccurate view of cheating behavior, which our study addressed effectively.

6.3. Interpreting the impacts of interventions from a user-centered perspective

Besides achieving the objective of cheating prevention, it was also important to evaluate different interventions from a user-centered perspective to infer their acceptability among test-takers. Hence, our study evaluated the interventions studying their impact on user experience and test-taking self-efficacy.

6.3.1. Interventions on user experience (UX)

Overall, the interventions did not significantly affect the test-taking experience. Note that our objective was not to improve UX in test-taking but rather to put a check on any unintended negative experiences possibly caused by the interventions. During their interaction with the test design, participants rated usability aspect (pragmatic UX) higher than emotional aspect (hedonic UX) of the experience. Positive emotions were more frequent than negative. In addition to online questionnaires, incorporating open-ended questions was necessary to gain a comprehensive understanding of participants' evaluations. This was because the possibility of emotional reactions stemming from their perceived test performance had the potential to impact their overall experiences.

The open-ended questions, capturing participants' subjective opinions on interventions, revealed that delayed intervention messages may distract test-takers, causing stress and impacting the test-taking experience (Sefcik et al., 2022). Fear of false accusations of cheating was more prominent in warning and monitoring interventions. This was expected because both these interventions aim at elevating alertness against cheating while honor code reminders persuade them not to cheat by activating moral reasonings. Overall, users had a balanced and neutral to positive experience, with only around 3% extreme emotive instances recorded.

6.3.2. Interventions on test-taking self-efficacy

Previous studies (Daffin & Jones, 2018; Hylton et al., 2016; Pleasants et al., 2022) have observed that anti-cheating interventions may affect test performance. We instead evaluated different interventions based on test-taking self-efficacy, which is a reliable proxy for test performance (Zimmerman, 1995). This was because participants were not given sufficient time to prepare for the test and were only assessed based on their beliefs of performing well. Therefore, test-taking self-efficacy was considered a valid user-centered indicator along with UX to evaluate the interventions. Similar to UX, test-taking self-efficacy showed no significant differences between groups with and without interventions, indicating no noticeable harm. There was a positive and moderate correlation (Table A.2 in Appendix) between test-taking self-efficacy and UX, suggesting a potential influence between participants' perceptions of self-efficacy and their test-taking experiences. However, since empirical justification for this relationship is lacking in the existing literature, we have reported both aspects separately to comprehensively evaluate the interventions from a user-centered perspective.

To summarize, cheating prevention is challenging in a BYOD (*Bring Your Own Device*) setup with more opportunity to cheat. This study points towards a potential direction for reducing the likelihood of engaging in cheating by effectively conveying anti-cheating information and reinstating test-takers' moral principles during an examination. Otherwise, test-takers tend to circumvent the test rules through alternative means (e.g., switching browsers or devices) to avoid detection. The results also infer that the delayed appearance of an intervention message may have impacted the perception of being watched. It supports the argument that an increased certainty of being caught also decreases the likelihood of cheating (Freiburger et al., 2017; Miller et al., 2011). However, such delayed appearance might be distracting, possibly causing negative emotions among the test-takers. The interventions likely do not affect test-taking experience or self-efficacy. The findings have relevant implications for the use of privacy-non-invasive anti-cheating mechanisms and their acceptability in the future. Finally, as there was no such real-time tracking of participants' activities, such a 'bluff' might curb cheating in the short run, but it is not likely to yield long-term results (Pleasants et al., 2022). We discuss the possible workarounds in section 7.

7. Actionable recommendations for practitioners and researchers

Our study's results aid educational and testing organizations concerned with privacy-non-invasive, unproctored online assessments. We will outline intervention recommendations, with the aim of guiding researchers and practitioners in further exploring and addressing the issue of reducing cheating.

- *Type of intervention*
  1. Honor code reminder: Presenting a honor code both before and during a test holds the potential for reducing cheating without negatively impacting user experience or self-efficacy. It is important to carefully craft the message to effectively engage moral reasoning. However, it is advisable to keep the reminder concise to prevent any potential distractions.
  2. Warning message: Employing a warning system to deter cheating could be another effective anti-cheating mechanism. However, it is important to note that if harsh penalties are explicitly mentioned, they may have a negative impact on performance, potentially stemming from the fear of false accusations.
  3. Monitoring message: Relying solely on monitoring messages may not be an adequate method to effectively prevent cheating. Our study highlights the need for additional measures to address potential issues such as browser or device switching.
- *Presentation of intervention*: Our study provides empirical evidence that the delayed implementation of interventions generates a sense of being monitored, irrespective of the specific intervention type. Therefore, honor code reminders and warning messages have the potential to provide more anti-cheating information compared to monitoring alone. However, it is worth noting that these interventions may also result in distraction and induce stress.
- *Reliability of intervention*: It is suggested that clear cheating detection criteria should be communicated to test-takers before the test in order to alleviate concerns about potential false accusations. Additionally, test policies should incorporate the consequences of cheating to ensure transparency and avoid any perception of insincerity. Moreover, slight modifications to the design of interventions over time may be necessary to prevent desensitization, as highlighted by Corrigan-Gibbs et al. (2015).

8. Limitation and future scope

Our study has some limitations that could be addressed in future works.

First, our study aimed to involve a more diverse range of participants beyond just university students, as assessment and exam situations are frequent also outside the University. It is recommended that future studies focus on including non-university students and analyzing how demographic variables can affect the prevention of cheating. Second, our study used a sample selected anonymously from Prolific, a crowd-working platform. While this method is cost-effective and fast, it has drawbacks such as higher selection bias and social responsibility bias compared to lab-based studies, as noted by Chandler et al. (2019) and Behrend et al. (2011). Third, future research could replicate the study design solely on student populations, as they are often a primary assessment target. Fourth, repeated experiments on a fixed sample over time could validate Chen et al. (2018)'s "learning to cheat" hypothesis that test-takers become comfortable cheating in unproctored tests. This would also show how cheating rates change over time.

Privacy-non-invasive interventions were chosen to avoid negative UX associated with privacy-invasive ones. However, we didn't compare the effectiveness of both techniques. Future research could address this comparison.

Next, self-efficacy is important in test-taking, as it impacts performance and the acceptability of test design. Although we measured self-efficacy in our study, we did not establish causal evidence or evaluate the reliability and validity of the modified self-efficacy index we adopted from Chen et al. (2001). Further research is needed to investigate these aspects.

To improve detection accuracy and avoid false positives/negatives, a cheating detection technique was used in a controlled experimental setting. However, the scope of variables in the field may not be entirely integrated. For example, replicating the detection method in an exam collaboration scenario could result in false negative errors. Future studies should use our findings and recommendations to produce additional validation in various field contexts. It would also be interesting to combine the tested cheating-preventive interventions with the aforementioned cheating-detection methods (e.g. entrapment used in Schultz et al. (2022), behavioral biometrics used in Pleasants et al. (2022) and response similarity used in Humbert et al. (2022)) in future studies to see whether privacy-non-invasive cheating prevention is achievable in a zero-tolerance policy.

Previous research (Jordan, 2001; Malesky et al., 2022) has shown a relationship between motivation to cheat and the likelihood of cheating, where intrinsically motivated individuals are less likely to cheat. We did not evaluate how pre-determination to cheat affects the effectiveness of anti-cheating interventions, so future research is needed. Our study's outcomes cannot be generalized as they may vary for different types of cheating practices and assessment types. This highlights the need for further research on different combinations.

Besides measuring UX and test-taking self-efficacy, technology acceptance models can gather additional insights into the acceptability of future test designs, with further consideration of creating an inclusive test design. The impact of interventions should be carefully validated in a real summative test scenario, where the stakes are high, and the possibility and types of cheating are higher, before any real-world application.

9. Conclusion

In high-stakes online summative assessments, cheating prevention is often a trade-off between their effectiveness and user experience. Online proctoring is effective but can invade privacy and negatively affect performance. Our research focuses on a user-centric solution by using anti-cheating interventions that do not invade privacy. We evaluated their effectiveness from a user-centered perspective.

We empirically compared three privacy-non-invasive anti-cheating interventions in a simulated online test with a diverse group of participants. The study captures the impact of the interventions on participants' cheating behavior and overall assessment experience from both quantitative and qualitative aspects. We used a unique cheating detection technique and found that implementing an anti-cheating intervention during a test proved successful in preventing cheating. Our experimentation involved three different privacy-non-invasive anti-cheating interventions: a honor code reminder, a warning message and a monitoring message. We found that interventions aimed at reinforcing test-takers' moral principles during tests, such as a reminder of a honor code, have the potential in reducing instances of cheating. Delayed interventions made test-takers feel watched and reduced cheating but could be distracting. None of the interventions affected overall assessment experience or self-efficacy. We recommend using these findings to design cheating-preventive unproctored online assessments. These results are promising and could be useful for future studies in real high-stakes online assessments.

CRediT authorship contribution statement

Suvadeep Mukherjee: Conceptualization, Software, Methodology, Validation, Investigation, Data curation, Writing – original draft, Project administration. **Björn Rohles:** Conceptualization, Methodology, Investigation, Supervision, Writing – review & editing. **Verena Distler:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Gabriele Lenzini:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition. **Vincent Koenig:** Conceptualization, Methodology, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Data availability

I have shared 2 data files in the 'Attach supplymentary data' step

Acknowledgements

We wish to thank Sophie Doublet for her invaluable assistance in designing the online test portal and conducting the data collection. We also acknowledge Margault Sacré for her kind support in reviewing the document. This research is the result of the project "Secure and Verifiable Electronic Testing and Assessment Systems" (INTER/ANR/20/14926102/SEVERITAS) funded by the Luxembourg National Research Fund (FNR) and the French National Research Agency (ANR).

Appendix A. Correlation Tables

Table A.1
Pairwise Correlations between conditions (as dummies) and age (continuous), respondent gender (male and female as dummies), university degree (Bachelors and higher degrees as dummy), and past experience with online tests (appeared in at least 1 test in last 2 years as dummy). Pearson's correlation coefficient. *p < 0.05

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1): Control	1.00								
(2): Honor code reminder	−0.33*	1.00							
(3): Warning	−0.34*	−0.33*	1.00						
(4): Monitoring	−0.33*	−0.32*	−0.33*	1.00					
(5): Male	−0.03	0.04	−0.03	0.02	1.00				
(6): Female	0.03	−0.05	0.03	−0.01	−0.98*	1.00			
(7): Age	−0.06*	−0.02	−0.01	0.08*	0.09*	−0.08	1.00		
(8): University degree	0.01	−0.04	0.09*	−0.05	−0.04	0.04	−0.14*	1.00	
(9): Past experience	0.05	−0.03	−0.01	−0.01	0.07*	−0.07*	−0.11*	−0.02	1.00

Table A.2
Pairwise Correlations between dependent variables. Pearson's correlation coefficient. *p < 0.05

	(1)	(2)	(3)	(4)	(5)	(6)
(1): Cheating behavior	1.00					
(2): UX-Pragmatic	0.11*	1.00				
(3): UX-Hedonic	−0.02	0.1*	1.00			
(4): Positive emotion	0.12*	0.18*	0.49*	1.00		
(5): Negative emotion	−0.03	−0.25*	−0.07*	0.02	1.00	
(6): Self-efficacy of test-taking	0.22*	0.36*	0.23*	0.38*	−0.26*	1.00

Appendix B. Materials of the Study Design

Table B 1
Full set of test questions along with the range of realistic values generated by the publicly available 'trap' websites for each of them

Question number	Test questions	Range of realistic answers generated by 'trap' websites
1	Xiaowei is double Andrea's age. Xiaowei will be triple Sofia's age in 2 years. If Sofia is 14 years old currently, what will be the summation of their ages (all three) after 20 years?	[101, 180]
2	In Thanksgiving, Mr Larry Bowden decided to distribute some candies among his friends in a manner that he gives half of his total stock of candies and one extra to the first friend, and then the half of the remaining stock along with one extra to the second and continues giving away in this fashion. His total stock exhausted after he gave away to 5 friends. How many candies were there in his stock initially?	[47, 500]
3	In a secret messaging communication, the word DAF is coded as 530 in a specific code language. In the same language, what will HIG be coded as?	[613, 999]
4	Vicky and Rosetta finally started their long planned farming business. Vincent invested $4332 for 18 months and Rosetta invested $3000 for 13 months. If both of them earned a profit of $2500 after 18 months then what is the share of Vincent in that profit? (calculate the value in $)	[1461, 1800]
5	Alina and Luce can do a piece of work in 36 days and 81 days respectively. Both worked for 5 days together and then Alina left. How many days will Luce take to finish the remaining work?	[37, 76]

Table B.2
The following open-ended questions were qualitatively evaluated only by the participants from treatment groups (honor code reminder, warning and monitoring conditions)

Open Question 1	The message was about......
Open Question 2	I think the message wanted to.......
Open Question 3	The message impacted my behavior in terms of.......
Open Question 4	The message made me feel......
Open Question 5	I think that the message makes other participants.......

Table B.3
Self-efficacy statements were adopted from Chen et al. (2001)'s scale. Only the bolded phrase ("in a similar test system") was appended for each item; Cronbach's $\alpha = 0.95$

Item 1	I will be able to achieve most of the goals that I have set for myself **in a similar test system**
Item 2	When facing difficult tasks, I am certain that I will accomplish them **in a similar test system**
Item 3	I think I can obtain outcomes that are important to me **in a similar test system**
Item 4	I believe I can succeed at most any endeavor to which I set my mind **in a similar test system**
Item 5	I will be able to successfully overcome many challenges **in a similar test system**
Item 6	I am confident that I can perform effectively on many different problems **in a similar test system**
Item 7	Compared to other people, I can do most tasks very well **in a similar test system**
Item 8	Even when things are tough, I can perform quite well **in a similar test system**

Appendix C. Pilot Studies

We carried out five pilot studies to assess various test-related characteristics. In the first study, we conducted eight usability tests with semi-structured interviews to prevent usability issues from affecting participants' test performance. We improved the design based on feedback from participants, and subsequent pilot studies focused on experimental procedures. We selected five aptitude questions based on participant response time and success rate and allocated 10 min for the final test. To address cheating, we pre-tested a detection procedure using trap websites. We also tested performance-based incentives and found that £15 was sufficient. Finally, we decided to display interventions after a 10-s delay to direct participants' attention to the information provided. This decision was based on Theeuwes (1992) and Theeuwes et al. (1998), as some participants ignored additional information during pilot studies due to high cognitive demands for problem-solving.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2023.104925.

References

Amigud, A., & Lancaster, T. (2019). 246 reasons to cheat: An analysis of students' reasons for seeking to outsource academic work. *Computers & Education, 134*, 98–107.

Balash, D. G., Kim, D., Shaibekova, D., Fainchtein, R. A., Sherr, M., & Aviv, A. J. (2021). Examining the examiners: Students' privacy and security perceptions of online proctoring services. In *Seventeenth symposium on useable privacy and security* (pp. 633–652). SOUPS 2021.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. englewood cliffs, nj: Princeton hall.

Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues, 46*(1), 27–46.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117–148.

Bandura, A., Freeman, W. H., & Lightsey, R. (1999). *Self-efficacy: The exercise of control*.

Barnhardt, B. (2016). The "epidemic" of cheating depends on its definition: A critique of inferring the moral quality of "cheating in any form". *Ethics & Behavior, 26*(4), 330–343.

Baume, M. (2019). Online proctored exams: Where and how are they used. In *Basics, practical scenarios and technical solutions for online proctoring at european universities and educational institutions* (pp. 5216–5225). 13th International Technology, Education and Development Conference.

Baumeister, R. F., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological Inquiry, 7*(1), 1–15.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*(3), 800–813.

Bing, M. N., Davison, H. K., Vitell, S. J., Ammeter, A. P., Garner, B. L., & Novicevic, M. M. (2012). An experimental investigation of an interactive model of academic cheating among business school students. *The Academy of Management Learning and Education, 11*(1), 28–48.

Brimble, M. (2016). *Why students cheat: An exploration of the motivators of student academic dishonesty in higher education*. Handbook of academic integrity.

Brinthaupt, T. M. (2004). Providing a realistic course preview to students. *Teaching of Psychology, 31*(2), 104–106.

Buck, R., & Davis, W. A. (2010). Marketing risk: Emotional appeals can promote the mindless acceptance of risk. In *Emotions and risky technologies* (pp. 61–80). Springer.

Buck, R., & Ferrer, R. (2012). Emotion, warnings, and the ethics of risk communication. In *Handbook of risk theory* (pp. 694–723).

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers & Education, 159*, Article 104024.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond mechanical turk. *Behavior Research Methods, 51*(5), 2022–2038.

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4*(1), 62–83.

Chen, B., West, M., & Zilles, C. (2018). How much randomization is needed to deter collaborative cheating on asynchronous exams?. In *Proceedings of the fifth annual ACM conference on learning at scale* (pp. 1–10).

Chin, E. C., Williams, M. W., Taylor, J. E., & Harvey, S. T. (2017). The influence of negative affect on test anxiety and academic performance: An examination of the tripartite model of emotions. *Learning and Individual Differences, 54*, 1–8.

Cohney, S., Teixeira, R., Kohlbrenner, A., Narayanan, A., Kshirsagar, M., Shvartzshnaider, Y., & Sanfilippo, M. (2021). Virtual classrooms and real harms: Remote learning at {US}. universities. In *Seventeenth symposium on useable privacy and security* (pp. 653–674). SOUPS 2021).

Conijn, R., Kleingeld, A., Matzat, U., & Snijders, C. (2022). The fear of big brother: The potential negative side-effects of proctored exams. *Journal of Computer Assisted Learning, 38*(6), 1521–1534.

Coohey, C., & Cummings, S. P. (2019). Evaluation of an online group intervention to improve test-taking self-efficacy and reduce licensure test anxiety. *Journal of Social Work Education, 55*(2), 376–388.

Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction, 22*(6), 1–23.

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the motivated strategies for learning questionnaire. *Learning and Individual Differences, 21*(4), 337–346.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

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

Dawson, P., & Sutherland-Smith, W. (2019). Can training improve marker accuracy at detecting contract cheating? A multi-disciplinary pre-post study. *Assessment & Evaluation in Higher Education, 44*(5), 715–725.

Doherty, S., & Nelson, R. (2010). Using projective techniques to tap into consumers' feelings, perceptions and attitudes… getting an honest opinion. *International Journal of Consumer Studies, 34*(4), 400–404.

Donoghue, S. (2000). Projective techniques in consumer research. *Journal of Consumer Sciences, 28*.

Eden, D. (1988). Pygmalion, goal setting, and expectancy: Compatible ways to boost productivity. *Academy of Management Review, 13*(4), 639–652.

Eden, D. (1996). From self-efficacy to means efficacy: Internal and external sources of general and specific efficacy. In *56th annual meeting of the Academy of Management*. Cincinnati, OH.

Eden, D., & Zuk, Y. (1995). Seasickness as a self-fulfilling prophecy: Raising self-efficacy to boost performance at sea. *Journal of Applied Psychology, 80*(5), 628.

Fazio, R. H., & Cooper, J. (1983). Arousal in the dissonance process. *Social psychophysiology: A sourcebook*, 122–152.

Fontaine, S., Frenette, E., & Hébert, M.-H. (2020). Exam cheating among quebec's preservice teachers: The influencing factors. *International Journal for Educational Integrity, 16*(1), 1–18.

Freiburger, T. L., Romain, D. M., Randol, B. M., & Marcum, C. D. (2017). Cheating behaviors among undergraduate college students: Results from a factorial survey. *Journal of Criminal Justice Education, 28*(2), 222–247.

Ghanem, C. M., & Mozahem, N. A. (2019). A study of cheating beliefs, engagement, and perception–the case of business and engineering students. *Journal of Academic Ethics, 17*, 291–312.

Harper, M. G. (2006). High tech cheating. *Nurse Education Today, 26*(8), 672–679.

Harper, R., Bretag, T., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: A survey of australian university staff. *Studies in Higher Education, 44*(11), 1857–1873.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction, 13*(4), 481–499.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In *Mensch & computer 2003* (pp. 187–196). Springer.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology, 25*(2), 91–97.

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*, 63–84.

Hu, S., Jia, X., & Fu, Y. (2018). Research on abnormal behavior detection of online examination based on image information. In *2018 10th international conference on intelligent human-machine systems and cybernetics (IHMSC)* (Vol. 2, pp. 88–91). IEEE.

Humbert, M., Lambin, X., & Villard, E. (2022). The role of prior warnings when cheating is easy and punishment is credible. *Information Economics and Policy, 58*, Article 100959.

Hylton, K., Levy, Y., & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education, 92*, 53–63.

Jordan, A. E. (2001). College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy. *Ethics & Behavior, 11*(3), 233–247.

Kam, C. C. S., Hue, M. T., & Cheung, H. Y. (2018). Academic dishonesty among Hong Kong secondary school students: Application of theory of planned behaviour. *Educational Psychology, 38*(7), 945–963.

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

Klass, E. T. (1978). Psychological effects of immoral actions: The experimental evidence. *Psychological Bulletin, 85*(4), 756.

Kujala, S., Walsh, T., Nurkka, P., & Crisan, M. (2014). Sentence completion for understanding users and evaluating user experience. *Interacting with Computers, 26*(3), 238–255.

Ladyshewsky, R. K. (2015). Post-graduate student performance in 'supervised in-class' vs.'unsupervised online'multiple choice tests: Implications for cheating and test security. *Assessment & Evaluation in Higher Education, 40*(7), 883–897.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group* (pp. 63–76). Springer.

Lerner, M. J. (1977). The justice motive: Some hypotheses as to its origins and forms 1. *Journal of Personality, 45*(1), 1–52.

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

Li, H., Xu, M., Wang, Y., Wei, H., & Qu, H. (2021). A visual analytics approach to facilitate the proctoring of online exams. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–17).

Lown, J. M. (2011). Development and validation of a financial self-efficacy scale. *Journal of Financial Counseling and Planning, 22*(2), 54.

Malesky, A., Grist, C., Poovey, K., & Dennis, N. (2022). The effects of peer influence, honor codes, and personality traits on cheating behavior in a university setting. *Ethics & Behavior, 32*(1), 12–21.

Manoharan, S. (2019). Cheat-resistant multiple-choice examinations using personalization. *Computers & Education, 130*, 139–151.

McCabe, D. L. (1993). Faculty responses to academic dishonesty: The influence of student honor codes. *Research in Higher Education, 34*(5), 647–658.

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education, 64*(5), 522–538.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Dishonesty in academic environments: The influence of peer reporting requirements. *The Journal of Higher Education, 72*(1), 29–45.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction, 3*(CSCW), 1–23.

Miller, A., Shoptaugh, C., & Wooldridge, J. (2011). Reasons not to cheat, academic-integrity responsibility, and frequency of cheating. *The Journal of Experimental Education, 79*(2), 169–184.

Milone, A. S., Cortese, A. M., Balestrieri, R. L., & Pittenger, A. L. (2017). The impact of proctored online exams on the educational experience. *Currents in Pharmacy Teaching and Learning, 9*(1), 108–114.

Minge, M., & Riedel, L. (2013). mecue-ein modularer fragebogen zur erfassung des nutzungserlebens. In *Mensch & computer* (pp. 89–98).

Mitchell, M. S., Baer, M. D., Ambrose, M. L., Folger, R., & Palmer, N. F. (2018). Cheating under pressure: A self-protection model of workplace cheating behavior. *Journal of Applied Psychology, 103*(1), 54.

Mlekus, L., Bentler, D., Paruzel, A., Kato-Beiderwieden, A.-L., & Maier, G. W. (2020). How to raise technology acceptance: User experience characteristics as technology-inherent determinants. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO), 51*(3), 273–283.

Moten, J., Jr., Fitterer, A., Brazier, E., Leonard, J., & Brown, A. (2013). Examining online college cyber cheating methods and prevention measures. *Electronic Journal of e-Learning, 11*(2), pp139–146.

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology, 41*(1), 167–194.

Park, S., & Avery, E. J. (2019). Development and validation of a crisis self-efficacy index. *Journal of Contingencies and Crisis Management, 27*(3), 247–256.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662.

Pleasants, J., Pleasants, J. M., & Pleasants, B. P. (2022). Cheating on unproctored online exams: Prevalence, mitigation measures, and effects on exam performance. *Online Learning, 26*(1).

Ranger, J., Schmidt, N., & Wolgast, A. (2020). The detection of cheating on e-exams in higher education—the performance of several old and some new indicators. *Frontiers in Psychology, 11*, Article 568825.

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

Schrepp, D. M. (2019). *User experience questionnaire handbook* (8 edition).

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(6), 103–108.

Schultz, M., Lim, K. F., Goh, Y. K., & Callahan, D. L. (2022). *Ok google: what's the answer? Characteristics of students who searched the internet during an online chemistry examination*. Assessment & Evaluation in Higher Education.

Schultz, M., Schultz, J., & Round, G. (2008). *Online non-proctored testing and its affect on final course grades* (Vol. 9, pp. 11–16). Cambridge: Business Rev.

Schwartz, B. M., Tatum, H. E., & Hageman, M. C. (2013). College students' perceptions of and responses to cheating at traditional, modified, and non-honor system institutions. *Ethics & Behavior, 23*(6), 463–476.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (Vol. 35, p. 37).

Schwarzer, R., & Jerusalem, M. (2010). The general self-efficacy scale (gse). *Anxiety, Stress & Coping, 12*(1), 329–345.

Sefcik, L., Veeran-Colton, T., Baird, M., Price, C., & Steyn, S. (2022). An examination of student user experience (ux) and perceptions of remote invigilation during online assessment. *Australasian Journal of Educational Technology, 38*(2), 49–69.

Shaffer, D. R. (1975). Some effects of consonant and dissonant attitudinal advocacy on initial attitude salience and attitude change. *Journal of Personality and Social Psychology, 32*(1), 160.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports, 51*(2), 663–671.

Shu, L. L., & Gino, F. (2012). Sweeping dishonesty under the rug: How unethical actions lead to forgetting of moral rules. *Journal of Personality and Social Psychology, 102*(6), 1164.

Stanley, K. D., & Murphy, M. R. (1997). A comparison of general self-efficacy with self-esteem. *Genetic, Social, and General Psychology Monographs, 123*(1), 79–100.

Thacker, E. J., et al. (2022). PhD thesis. In *Contract cheating and academic literacies: Exploring the landscape*. Keele University.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*(6), 599–606.

Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science, 9*(5), 379–385.

Truxillo, D. M., Bauer, T. N., & Sanchez, R. J. (2001). Multiple dimensions of procedural justice: Longitudinal effects on selection system fairness and test-taking self-efficacy. *International Journal of Selection and Assessment, 9*(4), 336–349.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Wuthisatian, R. (2020). Student exam performance in different proctored environments: Evidence from an online economics course. *International Review of Economics Education, 35*, Article 100196.

Yucas, A. (2015). *Chinese nationals charged with cheating by impersonation on us college tests*. The Guardian.

Zeidner, M. (1998). *Test anxiety: The state of the art*.

Zimmerman, B. J. (1995). Self-efficacy and educational development. *Self-efficacy in changing societies, 1*(1), 202–231.