The effect of performance goals and evaluation standard on cheating in an academic aptitude test –an experimental series

Tanja M. Fritz, Hernán González Cruz, Anja S. Göritz, Stefan Janke, Martin Daumiller

Angaben zur Veröffentlichung / Publication details:

Fritz, Tanja M., Hernán González Cruz, Anja S. Göritz, Stefan Janke, and Martin Daumiller. 2026. "The effect of performance goals and evaluation standard on cheating in an academic aptitude test –an experimental series." *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-026-09494-z.

Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-026-09494-z

The Effect of Performance Goals and Evaluation Standard on Cheating in an Academic Aptitude Test – An experimental series

Tanja M. Fritz[1] · Hernán González Cruz[2] · Anja S. Göritz[1] · Stefan Janke[2] · Martin Daumiller[3]

Received: 28 October 2025 / Accepted: 1 June 2026

Abstract
Relating students' achievement goals to their engagement in academic dishonesty has been a fruitful research avenue for understanding the motivational underpinnings of cheating behavior in academia. However, most study results are grounded in self-reported measures of academic dishonesty, and heterogeneous findings – particularly regarding the links between performance goals and academic dishonesty – warrant an experimental investigation into the conditions influencing these effects. Through three experiments, we examined the effects and interactions of performance goals (appearance-approach) and performance evaluation standards (result-based vs. process-based) on cheating behaviors in an academic aptitude test. In Study 1 (pilot, $N=146$), we tested the paradigm and developed suitable task types. In preregistered Study 2 ($N=238$), university students completed a supposed academic aptitude test in the lab using a 2 (appearance goal induction vs. no induction) × 2 (result-based vs. process-based evaluation) design, wherein cheating was measured by indicating having solved unsolvable tasks and using illicit means to solve knowledge questions. In preregistered Study 3 ($N=253$), we conducted a conceptual replication of Study 2 in an online setting. Taken together, the effects of goals and evaluation standards were inconsistent and differed between the type of cheating behavior and context. Methodological implications are discussed with respect to measures of cheating in both laboratory and online settings, while the practical implications focus on strategies to deter cheating through assessment modalities.

Keywords Performance goals · Evaluation standards · Academic dishonesty · Cheating · Experiment

Extended author information available on the last page of the article

Published online: 10 June 2026

1 Introduction

The exploration of factors influencing students' engagement in illicit academic behavior, such as cheating on tests or assignments, has been a longstanding subject in scientific inquiry (Marques et al., 2019). Academic cheating is a multifaceted phenomenon, impacting diverse domains, including the credibility of academic accreditations and the erosion of trust in educational institutions tasked with nurturing educated and ethical members of society (Resurreccion, 2012). Regarding factors that can motivate cheating, previous research has considered achievement goals (Anderman & Koenka, 2017): Students pursuing performance goals that emphasize competence demonstration are expected to be more inclined toward cheating, as they can prioritize outcomes over the means used to achieve them. However, mixed findings (see meta-analyses: Fritz et al., 2023; Krou et al., 2021) highlight contextual influences on this association.

In educational settings, written tests serve as a common context in which the relationship between achievement goals and cheating unfolds. A key contextual factor in this dynamic is the evaluation mode. The test format and criteria defining good performance (subsequently referred to as the *evaluation standard*) can be expected to shape the perceived ease and utility of cheating for goal attainment, potentially moderating the impact of performance goals on cheating. Understanding such moderators linked to the achievement task is crucial, as, unlike students' motivations, test formats provide a more easily adaptable factor to deter cheating. Such knowledge can thus contribute to the development of fair and valid testing modalities. Building on this premise, the present research investigated the interplay between performance goals and the evaluation standard, hypothesizing increased cheating among students pursuing appearance-approach goals (striving to appear competent), assessed solely based on correct answers rather than the solution process in a performance test.

2 The Interplay of Students' Achievement Goals and Evaluation Standards

Achievement goals are defined as future-directed cognitive representations of desired end states in achievement settings (Elliot et al., 2017). They are distinguished based on the standards defining success – competence development (mastery goals) and competence demonstration (performance goals) – leading to varied strategies for goal attainment. Due to its emphasis on performance outcomes over processes, the pursuit of performance goals is considered a contributing factor to academic dishonesty (Anderman & Koenka, 2017). To understand the mechanism behind this, one must consider the focus of these goals. Normative goals, one facet of performance goals, describe the striving to outperform others, which can be a driver to resort to dishonest means to gain competitive advantage. In scenarios where the ability to compare one's performance with peers is restricted (e.g., anonymized or private grade communication in conjunction with large course sizes minimizing student interaction), appearance goals, the second facet of performance goals, might manifest more prominently. They describe a striving to appear competent, a crucial element in the competitive

higher education environment where evaluations by senior academics influence academic and professional success. This competitive atmosphere makes appearance goals a dangerous catalyst for cheating to demonstrate good performance. We therefore focused on appearance goals for the research at hand.

However, the inconsistent findings of previous research suggest that the relationship between performance goals and academic dishonesty is not straightforward. The expectancy-value theory (Wigfield & Eccles, 2000), a prominent model on behavioral motivation, may offer insights in this regard. It posits that both the expectation for a successful outcome and the personal value of the outcome need to be high to motivate behavior. Accordingly, considering that strong performance goals set a high value on competence demonstration, their paving effect on academic dishonesty is predicated on the expectation that (undetected) cheating is possible and expedient. This expectation is shaped by task-specific factors, specifically the criteria set to assess and evaluate performance. Instructors can use different approaches to assess performance. One way is to merely evaluate whether students report correct answers (e.g., multiple-choice tests), another is to focus on solution strategies (e.g., giving points for correct steps towards a solution). These approaches differently influence the ease of cheating and the likelihood of being caught, both of which have been connected to the propensity of academic dishonesty (Bolin, 2004; Lonsdale, 2017; McCabe & Treviño, 1993; Yu et al., 2018). Under a result-based evaluation standard, cheating can be as simple as copying answers from another student, as identical responses are expected and not suspicious. In contrast, when students need to elaborate on their solution strategies (process-based evaluation standard), copying becomes more complex, takes longer, and similar answers on idiosyncratic solution pathways become more obvious. Qualitative data from middle grade student interviews support this notion (Zito & McQuillan, 2010), highlighting how task design and evaluation standards shape the motivation for and perceived ease of cheating.

In summary, while performance goals and a result-based evaluation standard both have the potential to incentivize students toward academic dishonesty, their combination could yield a potent mix of both a performance motive and opportunity (Schuchter & Levi, 2016), leading to increased cheating.

2.1 Research Aims and Hypotheses

The aim of this research was twofold: First, to examine the influence of performance goals (appearance-approach goals) on predicting cheating behavior in a performance test, and whether the evaluation standard for performance assessment moderates this relationship. To ensure conceptual precision, we focused on appearance-approach goal facet of performance goals, given their particular relevance in university contexts and prior evidence suggesting ambiguity in the effects of performance-avoidance goals on academic dishonesty (Fritz et al., 2023). For brevity, we refer to appearance-approach goals in the following when using the short term appearance goals. We hypothesized that focusing students on appearing competent and only correct answers mattering should stimulate cheating. Thus, our hypotheses were: (a) the induction of appearance-approach goals increases the likelihood of cheating behavior, (b) a result-based rather than process-based evaluation standard increases the likelihood of cheat-

ing behavior, and (c) that the effect of goals on cheating behavior is moderated by the evaluation standard in that cheating behavior is highest when appearance-approach goal induction and result-based evaluation standards are combined.

Second, this research builds on a laboratory study by Daumiller and Janke (2019), which found the expected interaction between appearance goal induction and a result-based evaluation standard, with highest cheating rates under these conditions. However, methodological limitations constrain the interpretability of these findings. In the process-based condition, participants were told that their solution strategies would be discussed afterward, which may have discouraged cheating due to the anticipated risk of detection. Moreover, the tasks did not directly implement a process-based response format but relied on the expectation of such a discussion. Consequently, the manipulation did not fully capture the conceptual meaning of process-based evaluation, which emphasizes elaboration, reasoning, and individual approaches within the task itself. The current research addresses these limitations by developing an improved paradigm in which the process-based evaluation condition was operationalized through tasks that required elaboration and a more direct recording of individual problem-solving strategies. This refinement strengthened the distinction between result-based and process-based conditions while maintaining comparable task cores across conditions. Study 1 served as a pilot to test and optimize this paradigm, Study 2 experimentally examined the effects of appearance goals and evaluation standards on cheating behavior, and Study 3 sought to replicate and extend these findings in an online context.

The hypotheses, research designs and statistical analysis plans of the two main studies were registered prior to data collection (Study 2: https://aspredicted.org/CPJ_HX4, Study 3: https://aspredicted.org/B6V_S1H). The data sets and analysis code are openly available at https://osf.io/tfxv3/?view_only=d8d4800f4a6b4cf7bac40e231266381d.

3 Study 1: pilot and paradigm development

Study 1 aimed to evaluate the feasibility and effectiveness of the refined experimental paradigm and to ensure that both evaluation standards elicited comparable engagement while maintaining distinct performance emphases. This pilot was conducted with $N = 146$ university students (58.9% women, 39.7% men, 1.4% diverse) and used the same experimental design as was planned for the main studies (see next section Main Studies Design and Statistical Analyses). We examined whether the cover story from the original study (development of new tasks for an academic aptitude test) remained plausible with the new tasks. As only 9.5% of participants suspected the study concerned matters of cheating or honesty, we kept the cover story for the subsequent studies. To assess the effectiveness of the evaluation standard manipulation, participants answered one item after each task type asking "*In your opinion*, *what is more important for a successful performance in these tasks?"* on a scale from 1 (*result*), over 4 (*both*), to 7 (*process*). To validate the various cheating opportunities provided, we used covert screen recording software to see how participants were solving the tasks.

3.1 Task development

Four task types were piloted. Each was selected to allow for result-based as well as process-based performance indicators and different indicators of cheating. Table 1 gives an overview of tasks, evaluation standard manipulations, and cheating opportunities.

Table 1 Overview of Tasks Tested in the Pilot (Study 1)

Task type	Description	Evaluation standard Participants were informed that the basis for their evaluation was…		Cheating opportunity
		Result-based	Process-based	
Difference spotting task	Finding differences between two images	Number of differences found	Search strategy (supposedly recorded by eye-tracking)	Over-reporting (only number of differences found needed to be given, but not pointed out)
Anagrams	Forming words from letter scrambles (multiple solutions possible with obvious and rare words)	Whether they found a correct solution word	Strategy used to find a word	Copying solution (the rare solution word was made visible when hovering over the anagram graphic)
Numbers puzzle	Rearranging numbers from 1 to 8 in a 3 × 3 matrix	Time and number of moves (fewer = better score)	Efficiency in moving numbers (supposedly recorded by the game site)	Under-reporting time/moves (information needed to be copied from the game site to the test)
Knowledge questions	Fact-based knowledge questions not easily solved by the average student	Whether they reported the correct solution	Explanation of thought process and topic knowledge	Internet search (wrong answers planted on a Q&A forum to identify cheaters)

Note. The anagrams and knowledge questions tasks were selected for the main studies and slightly adapted (see Procedure and Measures in the respective study)

3.2 Differences spotting task

Four image pairs, taken from Liu et al. (2021), were presented. Participants were asked to identify differences between the images without pointing them out, thereby allowing for overreporting. We examined whether claiming more differences than actually present (in the first two pairs) or presenting highly difficult tasks (i.e., image pairs that were in fact identical) would elicit cheating. Additionally, hovering the cursor over the images revealed a file name containing an exaggerated solution (e.g., "image1_5differences.jpg"). In the process condition, a fictitious eye-tracking procedure was introduced, and participants were informed that the software would track their visual search behavior, which would form the basis of their evaluation.

However, although overreporting was not rare (14.9%), only 2.0% of participants admitted intentional overreporting. This pattern suggests considerable measurement uncertainty (i.e., unintentional error), and participants did not discover the embedded solution feature. Furthermore, although the fictitious eye-tracking manipulation appeared credible to most participants, it seemed difficult to implement convincingly in later online studies on participants' own devices, where camera availability, privacy concerns, and visible webcam indicators could undermine plausibility. Participants also perceived the task as rather difficult, which was expected to result in longer completion times and potential disengagement. Finally, participants did not perceive a clear difference between the evaluation standards ($d = 0.24$, $p = .068$). Based on these considerations, this task type was not further considered.

3.3 Anagrams

Participants were presented with three anagrams with multiple possible solutions (e.g., ETIBR = German *breit*, *Brite*, *reibt*, *trieb*). The selected anagrams were based on a pre-study ($N = 11$), which identified a discernible pattern of commonly generated solutions and solutions that were rarely or never produced. As in the difference spotting task, the rare solution word was made visible when hovering the cursor over the graphic (e.g., "anagram1_solution.jpg"). The underlying rationale was to infer cheating based on the type of solution provided: commonly generated words would indicate independent problem-solving, whereas rare words would suggest copying from the embedded solution.

However, this feature was again rarely discovered by participants (3.4%). At the same time, 40.5% of participants reported the supposedly rare solution word in at least one of the three anagrams. Moreover, many participants indicated in a post-survey that they perceived the anagrams as relatively easy and saw little need to cheat. Taken together with the low rate of admitted cheating (5.5%), these findings suggest that the rare-solution paradigm lacks sufficient validity for reliably inferring cheating behavior, as the various solutions do not seem to differentiate in frequency as cleanly as expected. Nevertheless, the process condition, asking participants to elaborate on how they found a solution word, successfully induced a more process-based perception compared to the result-based condition ($d = 0.52$, $p = .002$). Therefore, we

retained the anagram task for further use but replaced the rare-solution approach with more established paradigms involving unsolvable anagrams (see descriptions in the main studies).

3.4 Numbers puzzle

Participants were directed via a link to an external website where they were required to rearrange numbers in ascending order within a 3×3 matrix. The site automatically recorded the number of moves and the time taken to complete the puzzle. After finishing, participants were asked to transfer this information to the survey, thereby creating an opportunity to misreport their performance. In the process condition, participants were informed that the website would track how they rearranged the numbers and that this information would form the basis of their evaluation. However, the manipulation appeared ineffective, as participants in the process condition perceived the evaluation standard as similarly result-based as those in the result condition (d = 0.04, p = .324), possibly due to a comparable focus on outcome metrics. Additionally, the puzzle was perceived as highly difficult, often resulting in prolonged solving times or failure to complete the task. Post-survey responses further indicated technical issues, such as unintentionally closing or restarting the puzzle, which introduced ambiguity regarding participants' actual performance. This uncertainty was reflected in inconsistent reporting behavior, with 14.7% of participants reporting lower, but also 11.5% reporting higher performance metrics than those recorded at task completion or termination. Taken together, these issues undermine the validity of the task as a measure of cheating behavior. Consequently, this task type was not retained for further use.

3.5 Knowledge questions

Participants were asked to answer two knowledge-based questions (e.g., "*What percentage of calcium is the human body made up of*?" and "*How many grapes are needed to produce a bottle of wine (0.75 L)*?"). These items were adapted from online knowledge quizzes and selected to be sufficiently difficult for the average student. To detect cheating, incorrect answers were deliberately planted on a prominent online Q&A platform (Quora), with the intention of identifying participants who searched for and reproduced these answers. This approach was designed to allow inference of cheating behavior without direct observation. However, this procedure could not be validated, as no participants were identified as using these answers or cheating in another way.

In the process condition, participants were required to articulate their reasoning and provide additional background knowledge related to the question. The manipulation of the evaluation standard was highly effective for this task type (d = 1.2, p < .001). For subsequent studies, we refined this approach by selecting more difficult, fact-based questions that are unlikely to be answered correctly without external assistance or guessing, rather than relying on the placement of incorrect answers on the internet.

3.6 Lessons from the pilot

We report on the piloted task types to inform future experimental cheating paradigms. For the purpose of our research aim, we selected the two task types that produced significant differences in the perception of the evaluation standard for the subsequent studies: anagrams and knowledge questions. Given the rather low cheating rates, we sought to revise these task types in ways that would increase both the opportunity for cheating and the validity of its detection. Specifically, we replaced the rare-solution anagram approach with a more established unsolvable-anagram paradigm, selected more difficult and less guessable fact-based knowledge questions, and allowed participants to use their own computers to reduce inhibition due to unfamiliar devices and increase the likelihood of internet searches. We also considered sample composition as a possible factor, as only half of the participants were promised additional monetary rewards for good performance, whereas the other half received guaranteed course credit.

4 Main studies design and statistical analyses

The studies' paradigm was approved by the ethics committee of the University of Mannheim (EK Mannheim 25/2018). Participants provided informed consent before participation. The target sample size for each main study was determined a priori based on simulation results for the expected interaction effects based on the prior study (Daumiller & Janke, 2019), assuming $\alpha = .05$ and power = .80. This resulted in a target sample of $N = 244$ ($n = 61$ per condition), and data collection continued until this minimum was reached.

Figure 1 shows an overview of the general experimental procedure. Like Study 1, both Study 2 (in-person experiment) and Study 3 (online replication) were implemented as a 2×2 between-subjects experiment with four conditions based on two independent variables: appearance-approach goals (goal induction vs. no goal induction) and evaluation standard (result-based vs. process-based). The two task types were used in both studies assessing cheating in the anagram task as lying (reporting having solved unsolvable anagrams) and use of illicit means in the knowledge questions (correct answers deemed highly indicative of cheating, as they are unlikely to be solved otherwise; see similar procedure for example in Zhao et al., 2023). Additionally, in Study 2, participants' behavior was observed by confederates. In Study 3, an anonymous online script recorded whether participants left the survey site, which was used as a proxy for illicit internet search, validated by correlating test performance with tab-switching events.

At the beginning of the test, participants were told that the study sought to validate tasks for an academic aptitude test and that they should imagine themselves in an assessment center situation for a job they would really like to have. To induce appearance-approach goals, in half of the cases, participants were also told that it is important for them to try to make a very good impression, and that they could receive up to €5 additional payment depending on their performance. All participants also

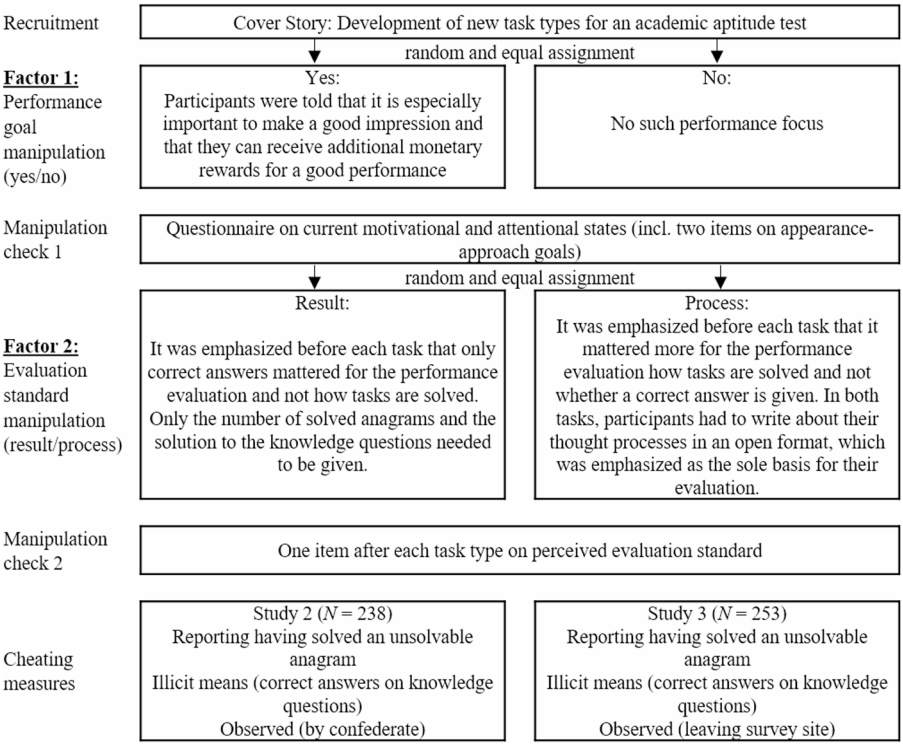


Fig. 1 Experimental Procedure of the Main Studies

had to note their main goal for the test in their own words, which was prompted to them again between the first and second task type (to strengthen the appearance goal induction).

To manipulate the evaluation standard, in the result-based condition, participants were introduced to the performance test with the statement that good results are most important for their performance evaluation and that it is less important how they solve the tasks. They only had to indicate whether they had solved an anagram or not and write down a solution to the knowledge questions. In the process conditions, participants were introduced to the test with the statement that their strategy in solving the tasks was most important for their performance evaluation and that it is less important, whether they solved a task, but how they solved it. In the anagram task, these participants were queried in an open-answer format how they approached the task and were reminded that this information was essential for their evaluation. Similarly, in the knowledge task, they had to write in an open-answer format about their thought process in answering the questions as the focus of their evaluation.

All analyses were carried out within the R environment (R Core Team, 2024; Version 4.1.2). The hypotheses were tested using logistic regression models for binary outcomes, reporting odds ratios (OR) as effect sizes and considering 95% confidence intervals when judging the uncertainty and statistical significance of effects.

5 Study 2: laboratory experiment

5.1 Participants

Seven cases were excluded from the data analysis: five participants indicated later in the test that they had already participated in the pilot study, and two participants indicated that technical issues required them to restart the test, which compromised the experimental manipulation. The final sample consisted of $N = 238$ German university students (67.0% women, 31.7% men, 1.3% diverse, $M_{Semesters} = 4.5$, $SD_{Semesters} = 2.4$). Students were approached on campus and invited to take part in the study. Most students (40.9%) were enrolled in teacher training programs, 11.3% in cultural studies and the humanities, 10.1% in linguistics and 7.5% in educational sciences. The remaining students were enrolled in diverse subjects including business administration, law, and engineering. We found no statistically significant differences between the experimental conditions regarding gender ($\chi^2 = 5.15(6)$, $p = .525$) and number of semesters ($F(3, 222) = 0.44$, $p = .723$), suggesting effective randomization. Anonymity of their responses was assured. Each participant was given €5 as compensation for their time.

5.2 Procedure and measures

Participants were randomly assigned to conditions. To ensure comparability within sessions, randomization accounted for completion times, as participants in the process-based evaluation groups had to provide longer responses. There were $n = 56$ participants in the no-goal/process group, $n = 66$ participants in the no-goal/result group, $n = 61$ participants in the goal/process group and $n = 55$ in the goal/result group. Between one and four students (group size $M = 2.4$, $SD = 1.0$) were recruited per session[1]. A confederate student was waiting in front of the test room and was introduced as another participant. The participants were then seated in the front row of the test room and the confederate in the second row to be able to observe the participants' screens. A second confederate was already seated in the second row as a supposed participant from the previous session who was still working on the test. Card boards between participants prohibited them from looking at each other's screens. Participants took the test on their own device. The experimenter emphasized that they should work on the test on their own and not use any other tools or means to solve the tasks. While participants were working on the first task (anagrams), the second confederate signaled that she had finished the test. The experimenter then opened the blackboard to show a link to a solution site (containing solutions for the knowledge questions task) and told the confederate (loud enough for all to hear) that she can compare her results with the solution. In that moment, the experimenters' phone was set to ring, she said she needed to take this

[1] We tested whether session groups had an influence on cheating outcomes in cumulative link mixed models and found only a variance of $SD = 0.2$ (lying in anagrams), 0.3 (cheating in knowledge questions), and 1.1 (observed cheating) in the log-odds for cheating between sessions, indicating no effect of session groups on cheating.

important call and left the room in a rush, leaving the link to the solution site visible to all. From there on, the two confederates started recording any illicit behavior by the participants. After everyone was finished, a confederate sent the experimenter a message via phone to reenter the room, who informed the participants (in addition to a written disclosure at the end of the test) about the actual study purpose and paid out their compensation.

As a manipulation check for the goal induction, before the performance test, participants answered two items on self-reported appearance-approach goals from the Academic Achievement Goal Questionnaire by Daumiller et al. (2019), mixed in among other motivational variables to not disclose the appearance-goal induction: "*In the following test*, *my goal is to be perceived as competent*" and "*In the following test*, *my goal is that it is noticed how good I am*", on a scale from 1 (*do not agree at all*) to 8 (*agree completely*). This part was introduced as it being important to know the preconditions under which the test is taken for the validation of the tasks, which aligned the appearance goal items well with the cover story. As a manipulation check for the evaluation standard manipulation, participants answered the same item as in the pilot after each task: "*In your opinion*, *what is more important for a successful performance in these tasks?"* on a scale from 1 (*result)* to 7 (*process*), with 4 indicating equal importance of both.

Based on pilot study findings, the anagram task was revised to a list of solvable and unsolvable anagrams (e.g., Liu et al., 2021; Lobel & Levanon, 1988; Wiltermuth, 2011), where participants had to indicate whether they found a solution or not. Participants were provided with a window in which they could rearrange the letters (emphasizing that this will not be scored); in the process conditions, they were told to observe their thoughts while finding a word which will be queried after the tasks as the basis of their evaluation. We added two more difficult fact-based questions to the piloted questions ("*When the liquid runs down the spout on the body of the jug*, *this is referred to as…?*" and "*What do you call the corps leader who leads the musicians in a marching band?*") to increase cheating opportunities.

Cheating was measured as lying when participants indicated having solved one or more unsolvable anagrams (three out of five anagrams), and as observed cheating (e.g., illicitly searching answers on the internet, using computer functions, exchanging with other participants, or checking the provided solution site) during the entire test after the experimenter had left the room. As stated in the preregistration, observed cheating was intended as the cheating measure for the second task type, however, the observing confederate students could not always discern at what stage of the test the target behavior occurred, and participants were also able to go back to previous tasks. We therefore tested as a third (not preregistered) dependent variable whether participants had the right answer to at least one of the knowledge questions (assuming that most students would not be able to solve them without other means). At the end of the test, participants were asked whether they had cheated on the tasks (this was not used as a dependent variable but served as comparison to the other measures).

5.3 Results

To ensure participants were taking the test carefully, we checked whether there were duration times lower than mean test time (22.9 min) minus two *SD* (16.9 min), which was not the case.

5.3.1 Manipulation checks

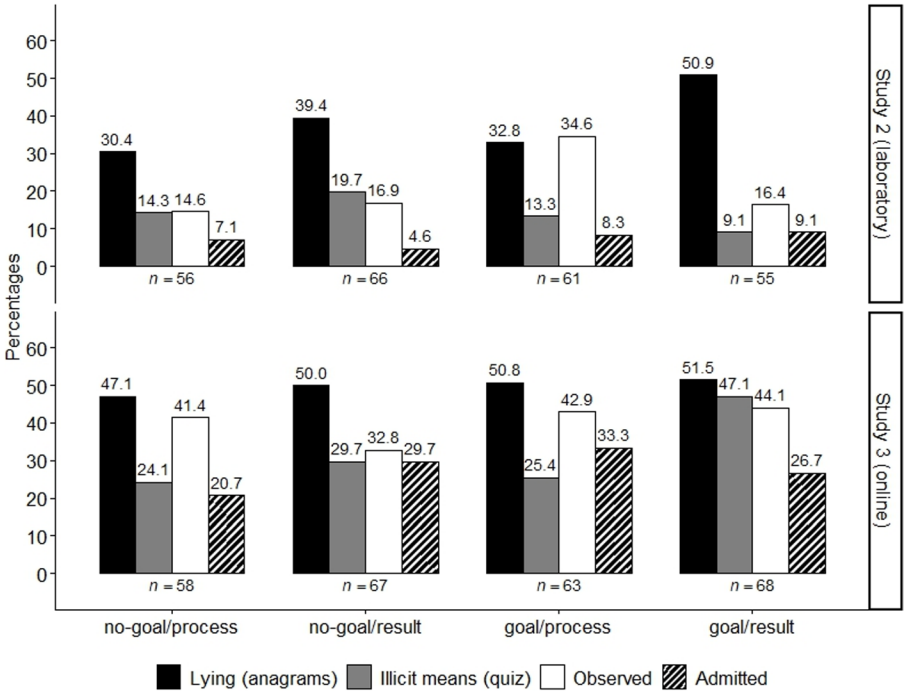
A *t*-test comparing the means of self-reported appearance-approach goals showed significantly higher means ($t=3.69$, one-sided $p < .001$, $d=0.48$) for participants receiving the appearance goal induction ($M=6.1$, $SD=1.3$) vs. no goal induction ($M=5.5$, $SD=1.4$). Because the goal induction included the prospect of additional performance-contingent payment, we also assessed extrinsic motivation to examine whether the manipulation reflected appearance-approach goals rather than a more general money-related motive ("*I only do this test for the money*", answers from 1 (*not at all true*) to 7 (*totally true*)) and found no significant differences in self-reported extrinsic motivation ($t \models 1.42$, one-sided $p = .157$, $d=0.18$) between the goal induction condition ($M=3.0$, $SD=1.8$) and the no goal induction condition ($M=3.4$, $SD=2.1$). Moreover, there were significant differences in the perception of the evaluation standard (anagrams: $t=2.69$, one-sided $p = .004$, $d=0.35$; knowledge questions: $t=7.53$, one-sided $p < .001$, $d=0.98$) between the process-based (anagrams: $M=4.0$, $SD=1.7$, knowledge questions: $M=4.1$, $SD=1.9$) and result-based (anagrams: $M=3.4$, $SD=1.8$, knowledge questions: $M=2.4$, $SD=1.6$) evaluation conditions. Therefore, we judged it appropriate to use the experimental condition variables as predictors in the models (we conducted robustness checks using the continuous variables, i.e., appearance-approach mean and evaluation standard variable, which did not change the results).

5.3.2 Cheating rates

The percentages of cheating rates in the different cheating variables per condition are in Fig. 2. In total, 38.4% of participants reported having solved at least one unsolvable anagram. In the knowledge questions, 14.1% of participants reported the right solution to at least one of the four knowledge questions. Confederates observed illicit behavior in 20.6% of participants. A total of 7.2% of participants admitted having cheated on the test.

5.3.3 Influence of experimental manipulations

We first tested main effects of the experimental conditions in logistic regressions, using the goal induction and evaluation standard as binary predictor variables (see results of all regressions in Table 2). A significant main effect of the evaluation standard was found for lying in the anagram task and a significant main effect of goal induction for observed behavior (however, the margins of the odds' CI include no difference in odds). Regarding right answers to at least one of the four knowledge questions, neither goal induction nor evaluation standard were statistically significant predictors.



Note. *Lying* = reporting at least one unsolvable anagram as solved (Study 2) or reporting more than two consecutive solved anagrams (Study 3); Illicit means = at least one correct answer in the knowledge questions; Observed = confederate observation of cheating behavior (Study 2) or tab-switching event during difficult knowledge questions (Study 3); Admitted = admitted having cheated in either one task type

Fig. 2 Percentages of Cheating Rates in the Experimental Groups of Study 2 and Study 3

We then tested interaction effects between the experimental conditions for each cheating variable. Only for observed cheating behavior, a statistically significant interaction was observed between goal induction and evaluation standard (OR = 0.33, p = .044). However, the confidence interval for this interaction was wide (0.08 to 1.17) and included 1 (i.e., no effect), indicating considerable uncertainty about the size and robustness of the interaction.

5.4 Interim discussion of study 2

The evidence collected in this study stands somewhat in contrast to the findings by Daumiller and Janke (2019). As we found only weak evidence for the interaction effect in the current study, a reassessment of the experimental design is prompted: First, participants were rather compliant with the cover story (15.0% reporting they did not cheat because of not wanting to bias the research results). We suspected that the involvement of student research assistants in participant recruitment and test administration fostered a peer-group dynamic, potentially enhancing participants' conscientiousness and honesty. Given the low incidence

Table 2 Results of Logistic Regression Models on Effects of Experimental Conditions as Main Effect Models and Moderator Effect Models on Cheating Variables

	Study 2 (laboratory)			Study 3 (online)		
	OR	95% CI	$p_{one-sided}$	OR	95% CI	$p_{one-sided}$
Lying						
Main effects model						
Goal induction	1.36	[0.80, 2.31]	.131	1.02	[0.62, 1.67]	.469
Evaluation focus	1.78	[1.05, 3.06]	.017	0.99	[0.60, 1.62]	.483
Moderator model						
Goal induction	1.12	[0.51, 2.46]	.389	1.19	[0.58, 2.43]	.321
Evaluation focus	1.49	[0.71, 3.21]	.149	1.15	[0.56, 2.35]	.352
Goal x Evaluation	1.43	[0.49, 4.14]	.257	0.75	[0.28, 2.01]	.284
Illicit means						
Main effects model						
Goal induction	0.62	[0.29, 1.28]	.101	1.76	[1.03, 3.03]	.021
Evaluation focus	1.06	[0.51, 2.22]	.438	2.13	[1.25, 3.37]	.003
Moderator model						
Goal induction	0.92	[0.32, 2.70]	.441	1.07	[0.47, 2.47]	.436
Evaluation focus	1.47	[0.57, 4.01]	.216	1.33	[0.60, 3.01]	.246
Goal x Evaluation	0.44	[0.09, 1.99]	.147	2.35	[0.79, 7.05]	.063
Observed						
Main effects model						
Goal induction	1.79	[0.94, 3.48]	.041	1.32	[0.80, 2.19]	.142
Evaluation focus	0.63	[0.32, 1.20]	.079	0.86	[0.52, 1.43]	.286
Moderator model						
Goal induction	3.10	[1.25, 8.27]	.009	1.06	[0.52, 2.20]	.435
Evaluation focus	1.20	[0.45, 3.33]	.361	0.69	[0.33, 1.45]	.164
Goal x Evaluation	0.31	[0.08, 1.17]	.044	1.52	[0.55, 4.20]	.208

Note. For each cheating variable, the upper rows represent main effects of goal induction and evaluation focus; the lower rows include the interaction term (*Goal* × *Evaluation*) in a separate moderator model. Each block thus represents two logistic regression models per dependent variable. Goal induction: 0 = no, 1 = yes, evaluation focus: 0 = process, 1 = result; Lying = reporting at least one unsolvable anagram as solved (Study 2) or reporting more than two consecutively solved anagrams (Study 3); Illicit means = at least one correct answer in the knowledge questions; Observed = confederate observation of cheating behavior (Study 2) or tab-switching event during difficult knowledge questions (Study 3)

of cheating, we aimed to adjust the study design to create a more ecologically valid setting that better captures real-world cheating behavior. Second, since 8.7% of participants reported abstaining from cheating due to a genuine interest in assessing their academic aptitude, we concluded to explicitly state that participants will not receive feedback on their scoring. Third, within the anagram task window where participants could rearrange letters, we questioned whether individuals hesitated to admit having solved an anagram without forming a solution word in this window, even though they were informed that this information would not be scored. To this end, we next conducted an online replication of the experiment to foster a less personal atmosphere and to modify the task design to address these considerations.

6 Study 3: online experiment

6.1 Participants

The study was conducted with students from a Germany-wide panel sample and through advertisements in courses at the university. The final sample consisted of $N = 253$ university students (74.3% women, 24.1% men, 1.6% diverse, $M_{\text{Semesters}} = 4.9$, $SD_{\text{Semesters}} = 2.2$). Students were enrolled in teacher training programs (83.8%)[2], psychology and (applied) social sciences (4.4%), business and management (3.2%), and media and communication programs (3.2%). We found no statistically significant differences between the experimental conditions regarding gender ($\chi^2 = 4.71(6)$, $p = .581$) and number of semesters ($F(3, 249) = 0.38$, $p = .768$), suggesting successful randomization. Anonymity was assured and each participant was given a €10 Amazon voucher as compensation for their time.

6.2 Procedure and measures

The study used the same cover story as in Study 2, presenting the research as part of a task development process for an academic aptitude test. To prevent participants from using the study as a self-assessment tool, the invitation stated that no individual performance feedback would be provided. After the debriefing, participants received the €10 voucher, regardless of their performance.

Participants were again randomly assigned to four conditions: There were $n = 58$ in the no-goal/process group, $n = 64$ in the no-goal/result group, $n = 63$ in the goal/process group, and $n = 68$ in the goal/result group. The appearance-approach goal induction was the same as in Study 2. In the process conditions, participants were required to write down their solving strategies on the same page as the anagram list. This change aimed to enhance the focus on the problem-solving process after eliminating the letter-rearranging window. The evaluation standard manipulation for the knowledge items was the same as in Study 2.

To address the potential issues with the anagram task in Study 2, this task was slightly changed (analogous to Wiltermuth, 2011): A list of five anagrams was presented (ETIBR, REENG, RIPDENT, SEHINCE, EAHFHLCS) wherein the third anagram was not solvable (no possible German word). Participants were required to indicate the number of consecutive anagrams they could solve. Consequently, participants reporting a score higher than two were coded as lying. The knowledge task was also improved, in that we used validated general knowledge questions with established item difficulty (Steger et al., 2021) to ensure that the average student would struggle to solve them without other means. The set comprised eight items from various knowledge domains with four particularly challenging questions (only these were considered in the analyses) where again, cheating was coded if at least one item was solved correctly. To compensate for not being able to directly

[2] This oversampling was likely due to the advertisement of the study in the courses taught by the authors (mostly teacher training programs). Closer inspection of the information given by these participants showed a diverse range of school types and subjects.

observe participants' behavior in the online setting, we utilized an anonymous log data script (Diedenhofen & Musch, 2017). This script retains data privacy as it only records whether participants leave the tab of the survey site, without capturing other content of their activities. As in Study 2, we asked participants at the end of the test whether they had cheated.

6.3 Results

Again, to ensure participants were taking the test carefully, we checked whether there were duration times lower than mean test time (22.1 min) minus two *SD* (16.1 min), which was not the case.

6.3.1 Manipulation checks

A *t*-test showed that participants receiving the appearance goal induction did not report significantly higher appearance-approach goals ($M = 6.0$, $SD = 1.5$, $t = 1.47$, one-sided $p = .072$, $d = 0.19$) than participants not receiving the goal induction ($M = 5.7$, $SD = 1.5$) (see General Discussion on elaboration).[3] Again, we found no significant differences in self-reported extrinsic motivation ("*I only do this test for the money.*", $t \nvDash 0.41$, one-sided $p = .341$, $d = 0.05$) between the goal induction condition ($M = 3.9$, $SD = 2.0$) and the no goal induction condition ($M = 4.0$, $SD = 2.0$). There were significant differences in the perception of the evaluation standard (anagrams: $t = 3.50$, one-sided $p < .001$, $d = 0.44$; knowledge questions: $t = 5.90$, one-sided $p < .001$, $d = 0.74$) between the process-based (anagrams: $M = 4.4$, $SD = 1.7$, knowledge questions: $M = 3.7$, $SD = 2.2$) and result-based (anagrams: $M = 3.6$, $SD = 1.9$, knowledge questions: $M = 2.2$, $SD = 1.7$) evaluation conditions.

6.3.2 Cheating rates

The percentages of cheating rates in the different cheating variables per condition are in Fig. 2. In total, 48.6% of participants reported having solved the unsolvable anagram, and 33.2% of participants reported the right solution to at least one of the four difficult knowledge questions. In the anagram task, 10.7% of participants admitted to having cheated, while in the knowledge task, the percentage was 25.7%. Between 32.8% (no-goal induction and result-based evaluation) and 44.1% (goal induction and result-based evaluation) of participants left the survey during the difficult knowledge questions at least once, which was significantly associated with reporting at least one correct answer to these questions in all groups except the goal/process group (see Table 3 for all correlations), suggesting tab-switching events as a good proxy for cheating behavior.

[3] We re-ran the analyses excluding participants with lower than 35 s time spent on the goal induction page (estimated minimum reading time), which did not change the results. There was also no significant correlation between time spent on the induction page and reported appearance-approach goals.

Table 3 Correlations Between Tab-Switching Events and Performance in Difficult Knowledge Items

Condition	%	*r*	*p*
No-goal/process	41.4	.34	.008
No-goal/result	32.8	.27	.028
Goal/process	42.9	.16	.216
Goal/result	44.1	.57	<.001

6.3.3 Influence of experimental conditions

To test main and interaction effects, we again ran separate logistic regressions for each cheating measure using the two factors as binary predictors (Table 2). We found no statistically significant main effect of both the appearance goal induction and evaluation standard for lying in the anagram task and tab-switching events. Within the knowledge questions task, we found a statistically significant main effect of both factors. We found no statistically significant interaction effect between the appearance goal induction and evaluation standard on all cheating measures.

7 General discussion

To summarize our findings, we did not find a consistent pattern of significant effects for appearance-approach goals, evaluation standards, or their interaction on cheating behavior across all conditions. However, comparing the laboratory study with the online study revealed notable differences between settings and cheating indicators. Cheating was more frequent in the online Study 3 than in the in-person Study 2: Two-sample tests for equality of proportions showed significantly higher rates both for the anagram measure, $\chi^2(1) = 5.38$, $p = .020$, and the knowledge questions measure, $\chi^2(1) = 24.04$, $p < .001$ (observed cheating was not compared across studies because the operationalizations differed substantially). Participants also admitted cheating more frequently in the online study, $\chi^2(1) = 35.44$, $p < .001$, although admitted cheating remained descriptively lower than the behavioral indicators in both settings. Furthermore, within both studies, participants were more likely to report solving an unsolvable anagram than to provide at least one correct answer to the knowledge questions, as indicated by McNemar tests (Study 2: $\chi^2(1) = 32.33$, $p < .001$; Study 3: $\chi^2(1) = 14.89$, $p < .001$). This pattern suggests that the ease and visibility of cheating behaviors (e.g., overstating performance vs. using external aids) influence their prevalence.

Although the higher cheating rates in the online setting are consistent with the assumption that less socially controlled environments may facilitate cheating behavior, this interpretation requires caution. The study settings differed not only in context (laboratory vs. online), but also in procedural details, task implementation, and the effectiveness of the appearance goal manipulation, which was weaker in Study 3. Accordingly, the inconsistent findings are best understood as reflecting sensitivity to both contextual and methodological features, including differences in cheating indicators and their validity.

Taken together, these findings underscore the importance of operational clarity when interpreting heterogeneous cheating effects. It is also important to acknowledge

that the implications of our research – drawn from two experiments and a pilot study – remain inherently preliminary. We therefore discuss our findings in relation to condition, measure, and context below. In line with the secondary aim of our research, we provide a comprehensive description of our rationale and the insights gained from developing task types to measure cheating in an experimental setting. By doing so, we aim to inspire future research and contribute to the development of more robust and valid measures of cheating.

7.1 Methodological considerations for research

7.1.1 Experimental manipulations

The lack of significant differences in reported appearance-approach goals among participants in the online experiment highlights the challenges of effectively inducing such goals experimentally (Kauffman & Young, 2015; Özdemir, 2014), especially in an anonymous and private setting. While the observed mean differences were in the expected direction, the lack of statistical significance indicates that the manipulation may not have been strong enough to influence behavior. Appearance-approach goals are fundamentally tied to concerns about others' evaluations and the desire to appear competent in the eyes of an audience. In the absence of a perceived observer or social presence, the motivational salience of these goals may be diminished, that is, the extent to which the goal induction translates into behavior may be attenuated. To enhance the induction of performance goals in online settings (possibly also including a normative goal component), it may be beneficial to introduce elements that simulate social presence or external evaluation. For instance, incorporating features such as virtual peer comparisons, feedback from a fictitious evaluator, or real-time performance tracking visible to others could make the goal manipulation more salient. Additionally, emphasizing scenarios where participants believe their performance might be reviewed or shared could further align the online setting with the evaluative context that underpins appearance-approach goals.

The implications of no significant goal differences in Study 3, as indicated by the manipulation check, for significant or non-significant effects in the analyses are noteworthy. First, this could explain the difference in findings between Study 2 and 3, as well as in relation to the prior study (Daumiller & Janke, 2019). Second, the significant effects for illicit means could be interpreted in a way that the manipulation may have influenced cheating behavior through mechanisms other than the intended activation of appearance-approach goals. For instance, the manipulation could have inadvertently heightened general competitive tendencies or test-related anxiety, which may have contributed to increased cheating. Conversely, the non-significant effects on other cheating indicators could either reflect a genuine absence of goal-related effects or the said diminished translation of goal induction into actual behavior in the anonymous online setting. More generally, because the goal induction was combined with the prospect of performance-contingent payment, it cannot be ruled out that the manipulation also affected behavior through a more general motivational incentive rather than exclusively through appearance-approach goals.

Our manipulation check confirmed that the result-based evaluation standard was effective in shifting participants' perceptions, particularly in the knowledge task. In contrast, the process-based evaluation standard resulted in more ambiguous perceptions, with participants' responses centering around the midpoint of the scale, suggesting that they perceived a mix of both evaluation types. Accordingly, any observed effects of the evaluation standard should be interpreted cautiously, as the manipulation did not produce fully distinct perceptions of process- versus result-based evaluation. This may arise from the tasks in the process conditions incorporating both result and process formats. Designing tasks devoid of any result-based answer element proves challenging; even in the context of more process-oriented test formats at university, some form of answer usually must be generated. Furthermore, there is an inherent challenge in making some task formats more process-oriented, especially knowledge-based tasks, as participants' access to their reasoning might be limited (e.g., how do you come to know what the capital of a country is?). One way to strengthen the manipulation, especially in the direction of a process-oriented evaluation, would be to indicate concrete points given for steps in the process. In this approach, one could even design completely equal tasks but vary the allocation of points for the evaluation.

7.1.2 Cheating measures

Each cheating measure used in this study requires some degree of inference about participants' cheating behavior. For the anagram task, it is possible that participants genuinely believed they had solved an unsolvable anagram. Similarly, even when using validated difficult knowledge items, there is still a chance of some students being able to answer the questions genuinely. Although the observed correlations between correct answers and tab-switching in Study 3 supports the assumption that answers were likely obtained through internet searches, this cannot be conclusively verified. Furthermore, the tab-switching measure, due to preserving participants' anonymity, does not capture the content of browser activity. As a result, it remains unclear whether participants used the internet to look up answers, access ancillary information, or engage in unrelated activities. Additionally, the measure does not account for alternative forms of cheating, such as using a smartphone or consulting another person. These limitations underscore the inherent challenges in accurately detecting cheating behavior in experimental settings, especially in online contexts.

Another issue concerns the external validity of these cheating measures. The second aim of this study was to develop tasks more closely aligned with typical knowledge assessments conducted in university settings. This approach addresses a limitation in existing research, where cheating experiments often rely on tasks less relevant for academic assessment such as lying about coin flip outcomes (see, for example, Gerlach & Teodorescu, 2022). In contrast to such tasks based on random outcomes, our tasks involved more personal involvement, cognitive engagement, and knowledge retrieval. Nonetheless, future research should develop experimental paradigms that replicate real-world academic scenarios, such as solving exam-style problems with prior learning phases or completing timed assignments. Furthermore,

future research should continue to seek novel ways to conduct experimental field research in the context of high-stakes exam (e.g., Zhao et al., 2023), although the ethical challenges of inducing performance goals and manipulating evaluation standards are higher.

7.2 Implications for educational practice and future research

Our findings provide only limited and preliminary implications for educational practice, and any conclusions should be interpreted with caution. Additional research is needed to better understand the impact and interplay of evaluation standards and achievement goals on cheating behavior.

At the same time, the present results highlight the practical relevance of further investigating cheating behavior under different testing conditions. In particular, cheating rates were significantly higher in the online than in the laboratory setting, corroborating previous findings (e.g., Adzima, 2020; Janke et al., 2021). This suggests that more anonymous environments may facilitate cheating behavior by increasing opportunities while simultaneously decreasing perceived social repercussions (Singh et al., 2021; Steger et al., 2018). However, as discussed above, this difference cannot be attributed to contextual factors alone, given concurrent differences in study design and implementation. This underscores the need for future research to more systematically disentangle contextual and procedural influences on cheating.

More broadly, our findings contribute to ongoing discussions about the design of academic assessments. Prior research has suggested that simple recognition or knowledge-reproduction tests may be vulnerable to cheating and limited in promoting deeper learning (Roediger & Marsh, 2005; Simkin & Kuechler, 2005; Stagner-Hall, 2012). While the present results do not allow for strong conclusions about optimal test design, they reinforce the importance of continuing to investigate how different assessment formats interact with opportunities and motivations for cheating.

In an era where knowledge-based answers are readily accessible, particularly in digital learning environments, this line of research becomes increasingly important. Technological developments, including AI-based tools capable of generating sophisticated responses, further complicate the challenge of ensuring academic integrity (Evangelista, 2025). Future research should therefore build on the present findings to examine how assessment formats can be designed to both foster meaningful learning and reduce opportunities for cheating behavior. While technological control mechanisms such as proctoring may offer partial solutions (Pleasants et al., 2022), a multifaceted approach grounded in a better empirical understanding of cheating behavior will likely be necessary (Tight, 2024).

7.3 Limitations

Apart from the limitations of the tasks and derived cheating measures already discussed above, this research acknowledges limitations in the external validity inherent in experimental studies. A post-test survey exploring participants' reasons for refraining from cheating revealed two prevalent explanations: the per-

sonal insignificance of test results to participants and not wanting to distort the study's results. Conversely, in real-world settings like university examinations, students typically attach high personal significance to their performance, and the ramifications of cheating, such as academic penalties, are more personally impactful. These higher stakes likely introduce additional moral, motivational, and risk-related considerations that were not present in our experimental setup, where cheating had different consequences. While Cohn and Maréchal (2018) established significant associations between laboratory measures of cheating and rule-violating behavior in schools, it remains plausible that cheating in an actual university assessment involves different predictors and perceived consequences. Notably, the influence of performance goals in such contexts can be expected to surpass that in experimentally induced goals.

In addition, the generalizability of the findings is limited by the composition of the samples. With the exception of additional student panel participants in Study 3, participants were recruited from the same German university. Although they represented diverse fields of study, the findings may not generalize to students from other institutions, educational systems, or cultural contexts.

8 Conclusion

Knowledge of situational factors inherent in the testing environment is indispensable for developing assessments designed to deter academic dishonesty. Our findings highlight test design as a practical and controllable tool for educators, offering a more immediate and actionable intervention compared to the complex task of influencing students' achievement motivations. However, inconsistencies in our findings highlight the influence of context, task types, and other procedural factors on such levers. Advancing this line of research will require addressing the methodological challenge of reliably and validly inducing performance goals for experimental investigation, while simultaneously designing naturalistic test environments and tasks that reflect real-world academic settings.

Acknowledgements This research did receive funding: Martin Daumiller and Stefan Janke received funding from Deutsche Forschungsgemeinschaft (Grant ID DA 2392/1-1 | JA 3137/1-1).

Author contributions TMF: data curation, formal analysis, investigation, methodology, resources, software, visualization, writing – original draft, writing – review & editing; HGC: software, validation, writing – review & editing; ASG: data curation, resources, writing – review & editing; SJ: conceptualization, funding acquisition, supervision, project administration, writing – review & editing; MD: conceptualization, formal analysis, funding acquisition, methodology, project administration, supervision, writing – review & editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The hypotheses, research designs and statistical analysis plans of the two main studies were registered prior to data collection (Study 2: https://aspredicted.org/CPJ_HX4, Study 3: https://aspredicted.org/B6V_S1H). The data sets and analysis code are openly available at https://osf.io/tfxv3/?view_only=d8d4800f4a6b4cf7bac40e231266381d.

Declarations

Competing interests The authors declare no competing interests.

References

Adzima, K. (2020). Examining online cheating in higher education using traditional classroom cheating as a guide. *Electronic Journal of E-Learning*, *18*(6), 476–493. https://doi.org/10.34190/JEL.18.6.002

Anderman, E. M., & Koenka, A. C. (2017). The relation between academic motivation and cheating. *Theory Into Practice*, *56*(2), 95–102. https://doi.org/10.1080/00405841.2017.1308172

Bolin, A. U. (2004). Self-control, perceived opportunity, and attitudes as predictors of academic dishonesty. *The Journal of Psychology*, *138*(2), 101–114. https://doi.org/10.3200/JRLP.138.2.101-114

Cohn, A., & Maréchal, M. A. (2018). Laboratory measure of cheating predicts school misconduct. *The Economic Journal*, *128*(615), 2743–2754. https://doi.org/10.1111/ecoj.12572

Daumiller, M., & Janke, S. (2019). The impact of performance goals on cheating depends on how performance is evaluated. *AERA Open*, *5*(4), 2332858419894276. https://doi.org/10.1177/2332858419894276

Daumiller, M., Dickhäuser, O., & Dresel, M. (2019). University instructors' achievement goals for teaching. *Journal of Educational Psychology*, *111*(1), 131–148. https://doi.org/10.1037/edu0000271

Diedenhofen, B., & Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*, 1444–1459. https://doi.org/10.3758/s13428-016-0800-7

Elliot, A. J., Dweck, C. S., & Yeager, D. S. (Eds.). (2017). *Handbook of competence and motivation: Theory and application*. Guilford.

Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology*, *17*(1), ep559. https://doi.org/10.30935/cedtech/15775

Fritz, T., González Cruz, H., Janke, S., & Daumiller, M. (2023). Elucidating the associations between achievement goals and academic dishonesty: A meta-analysis. *Educational Psychology Review*, *35*(1), 33. https://doi.org/10.1007/s10648-023-09753-1

Gerlach, P., & Teodorescu, K. (2022). Measuring dishonest behavior: Hidden dimensions that matter. *Current Opinion in Psychology*, *47*, 101408. https://doi.org/10.1016/j.copsyc.2022.101408

Janke, S., Rudert, S. C., Petersen, Ä., Fritz, T. M., & Daumiller, M. (2021). Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity? *Computers and Education Open*, *2*, 100055. https://doi.org/10.1016/j.caeo.2021.100055

Kauffman, Y., & Young, M. F. (2015). Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance. *Computers & Education*, *83*, 44–56. https://doi.org/10.1016/j.compedu.2014.12.016

Krou, M. R., Fong, C. J., & Hoff, M. A. (2021). Achievement motivation and academic dishonesty: A meta-analytic investigation. *Educational Psychology Review*, *33*(2), 427–458. https://doi.org/10.1007/s10648-020-09557-7

Liu, J., Shen, Q., Zhang, J., Beyens, U., Cai, W., Decety, J., & Li, H. (2021). The Difference Spotting Task: A new nonverbal measure of cheating behavior. *Behavior Research Methods*, *53*(5), 1935–1944. https://doi.org/10.3758/s13428-020-01526-w

Lobel, T. E., & Levanon, I. (1988). Self-esteem, need for approval, and cheating behavior in children. *Journal of Educational Psychology*, *80*(1), 122–123. https://doi.org/10.1037/0022-0663.80.1.122

Lonsdale, D. (2017). Intentions to cheat: Ajzen's planned behavior and goal-related personality facets. *The Journal of Psychology*, *151*(2), 113–129. https://doi.org/10.1080/00223980.2016.1241737

Marques, T., Reis, N., & Gomes, J. (2019). A bibliometric study on academic dishonesty research. *Journal of Academic Ethics*, *17*(2), 169–191. https://doi.org/10.1007/s10805-019-09328-2

McCabe, D. L., & Treviño, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *Journal of Higher Education*, *64*(5), 522–538. https://doi.org/10.1080/00221546.1993.11778446

Özdemir, A. (2014). Students' achievement goals and underlying reasons: Their relation to intrinsic motivation and cheating [Master's thesis, Bilkent University]. http://hdl.handle.net/11693/15945

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), 268–284. https://doi.org/10.24059/olj.v26i1.2620

R Core Team (2024). R: A language and environment for statistical computing (Version 4.1.2). R Foundation for Statistical Computing. https://www.R-project.org/

Resurreccion, P. F. (2012). The impact of faculty, peers and integrity culture in the academe on academic misconduct among Filipino students. *International Journal of Academic Research in Business and Social Sciences*, *2*(12), 33–50. https://www.researchgate.net/publication/304169734

Roediger, I. I. I., H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Schuchter, A., & Levi, M. (2016). The fraud triangle revisited. *Security Journal*, *29*, 107–121. https://doi.org/10.1057/sj.2013.1

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73–98. https://doi.org/10.1111/j.1540-4609.2005.00053.x

Singh, R. K., Voggeser, B. J., & Göritz, A. S. (2021). Beholden: The emotional effects of having eye contact while breaking social norms. *Frontiers in Psychology*, *12*, 545268. https://doi.org/10.3389/fpsyg.2021.545268

Steger, D., Schroeders, U., & Gnambs, T. (2018). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, *36*(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Steger, D., Schroeders, U., & Wilhelm, O. (2021). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*, *28*(3), 1004–1017. https://doi.org/10.1177/10731911209149

Tight, M. (2024). Challenging cheating in higher education: A review of research and practice. Assessment &. *Evaluation in Higher Education*, *49*(7), 911–923. https://doi.org/10.1080/02602938.2023.2300104

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, *115*, 157–168. https://doi.org/10.1016/j.obhdp.2010.10.001

Yu, H., Glanzer, P. L., Johnson, B. R., Sriram, R., & Moore, B. (2018). Why college students cheat: A conceptual model of five factors. *The Review of Higher Education*, *41*(4), 549–576. https://doi.org/10.1353/rhe.2018.0025

Zhao, L., Peng, J., Yang, X., Yan, W., Ke, S., Dong, L. D., Li, Y., Ma, J., & Lee, K. (2023). Effects of honor code reminders on university students' cheating in unproctored exams: A double-blind randomized controlled field study. *Contemporary Educational Psychology*, *75*, 102213. https://doi.org/10.1016/j.cedpsych.2023.102213

Zito, N., & McQuillan, P. J. (2010). Cheating themselves out of an education: Assignments that promote higher-order thinking and honesty in the middle grades. *Middle School Journal*, *42*(2), 6–16. https://doi.org/10.1080/00940771.2010.11461752

Authors and Affiliations

Tanja M. Fritz[1] · Hernán González Cruz[2] · Anja S. Göritz[1] · Stefan Janke[2] · Martin Daumiller[3]

✉ Tanja M. Fritz
tanja.fritz@phil.uni-augsburg.de

Hernán González Cruz
hernan.gonzalez@uni-mannheim.de

Anja S. Göritz
anja.goeritz@uni-a.de

Stefan Janke
stefan.janke@uni-mannheim.de

Martin Daumiller
martin.daumiller@psychologie.uni-freiburg.de

[1] University of Augsburg, Augsburg, Germany

[2] University of Mannheim, Mannheim, Germany

[3] University of Freiburg, Freiburg, Germany