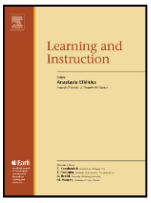
Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: http://ees.elsevier.com



People who cheat on tests accurately predict their performance on future tests

Monika Undorf [a,*], Eric Y. Mah [b], Dawn-Leah L. McDonald [c], Zachariah I. Hamzagic [c], Ryan Burnell [d], Maryanne Garry [d], Daniel M. Bernstein [c]

[a] Department of Psychology, School of Social Sciences, University of Mannheim, Germany
[b] Department of Psychology, University of Victoria, Canada
[c] Department of Psychology, Kwantlen Polytechnic University, Canada
[d] School of Psychology, University of Waikato, New Zealand

ARTICLE INFO

Keywords
Monitoring accuracy
Cheating
Academic dishonesty
Metacognition
Metamemory

ABSTRACT

Studies suggest that people who cheat on a test overestimate their performance on future tests. Given that erroneous monitoring of one's own cognitive processes impairs learning and memory, this study investigated whether cheating on a test would harm monitoring accuracy on future tests. Participants had the incentive and opportunity to cheat on one (Experiments 1, 2, and 3, with N = 90, 88, and 102, respectively) or two (Experiment 4, N = 214) of four general-knowledge tests. Cheating produced overconfidence in global-level performance predictions in Experiment 2 (Cohen's $d \geq$ 0.35) but not in Experiments 1 or 4. Also, cheating did not affect the absolute or relative accuracy of item-level performance predictions in Experiments 3 or 4. A Bayesian meta-analysis of all experiments provided evidence against cheating-induced overconfidence in global- and item-level predictions. Overall, our results demonstrate that people who cheat on tests accurately predict their performance on future tests.

1. Introduction

Cheating on tests is a pervasive problem in high school and college classrooms. Several large-scale studies have reported that more than 60% of high school students and more than 30% of college students admitted to test cheating (McCabe, Butterfield, & Trevino, 2012). Test cheating has significant adverse consequences. Cheating students may receive undeservedly high grades and unfair advantages over those who do not cheat (Bretrag, 2016). Cheating can also undermine attempts to foster students' ethical responsibility, and can damage institutional reputation (Passow, Mayhew, Finelli, Harding, & Carpenter, 2006). Also important, cheating has a detrimental impact on learning because students miss valuable learning opportunities (McDaniel, Anderson, Derbish, & Morrisette, 2007). For instance, Arnold (2016) reported that cheating on online practice tests was associated with low performance on final tests. Furthermore, cheating on tests prevents teachers from giving useful feedback to students and tailoring instruction to student needs (Passow et al., 2006).

Cheating's effects on metacognitive monitoring, however, have remained largely unexplored. Metacognitive monitoring refers to people's knowledge of their own cognitions, and particularly their own learning and memory processes (Dunlosky & Metcalfe, 2009; Koriat, 2007). There is ample evidence that accurate monitoring of one's knowledge fosters learning and memory, whereas poor monitoring impairs learning and memory. For instance, a meta-analysis showed that accurate monitoring positively predicts academic performance even when controlling for intelligence (Ohtani & Hisasaka, 2018). Other studies have differentiated between high monitoring effectiveness in terms of absolute and relative accuracy. *Absolute accuracy* pertains to whether people's metacognitive judgments correspond to their level of actual performance or, alternatively, whether people are over- or underconfident (for a different conceptualization, see Schraw, 2009). If a person predicts that she will recall six items at test and indeed remembers six items, then absolute accuracy is excellent. Consistent with the idea that high absolute accuracy supports learning, training undergraduates to accurately assess their own test performance improved their actual performance (Nietfeld, Cao, & Osborne, 2006). Dunlosky and Rawson (2012) experimentally increased the absolute accuracy with which people judged the correctness of recalled key term definitions. They found that high absolute accuracy of judgments improved final retention of definitions. Also, individual differences in absolute accuracy predicted final test performance, with high accuracy being associated with better retention.

* Corresponding author. Schloss, Ehrenhof-Ost, 68131, Mannheim, Germany.
E-mail address: undorf@uni-mannheim.de (M. Undorf)

https://doi.org/10.1016/j.learninstruc.2019.101295
Received 3 June 2019; Received in revised form 19 November 2019; Accepted 25 November 2019
Available online xxx
0959-4752/

The second aspect of metacognitive accuracy that is essential for learning and memory is *relative accuracy*, which assesses the extent to which metacognitive judgments distinguish between correct and incorrect responses. If a person predicts that she will recall one item but not another item at test, and indeed remembers only the former item, then relative accuracy for these items is excellent. Consistent with the idea that high relative accuracy benefits learning, relative accuracy of people's recall predictions from a first study-test cycle predicted their memory performance in three subsequent study-test cycles (Thiede, 1999). Structural equation modeling revealed that high relative accuracy of confidence judgments had a positive direct effect on memory for an educational film in 9- and 11-year old children (Roebers, Krebs, & Roderer, 2014). Together, these findings suggest that accurate metacognitive monitoring fosters learning and memory.

Cheating may harm monitoring accuracy and in turn impair people's learning and memory. Consider a student who has the questions that will appear on an upcoming exam. As a result of this information, the student does well on the exam. Will this student think of himself as an incompetent cheater or as a high achiever who just took a shortcut? A great deal of research suggests the latter. Psychologically healthy people have unrealistically positive self-evaluations (Alicke, 1985; Taylor & Brown, 1988). They interpret ambiguous information so as to confirm their positive self-views and ignore or rationalize negative information about themselves (Lord, Ross, & Lepper, 1979; Pyszczynski & Greenberg, 1987). If a cheating student attributes his good performance to high ability, however, this may result in him overestimating his knowledge and, consequently, underachieving on later exams. Indeed, people often expect that their future level of performance will match their current level of performance, even when there is good reason to expect differences (Critcher & Rosenzweig, 2014; Kornell & Hausman, 2017). For instance, memory predictions but not recall performance are largely unaffected by how many times people are told that they will be allowed to study in the future (Kornell & Bjork, 2009).

Most relevant to the current study, Chance, Norton, Gino, and Ariely (2011) found that cheating on a test inflated global performance predictions for future tests. Chance et al. (2011) had university students (Experiment 2: $N = 131$, Experiment 3: $N = 78$, Experiment 4: $N = 136$) complete a general-knowledge test with 10 items. The researchers gave half their participants answers to the test at the bottom of the page and instructions to use the answers only to mark their test. Not surprisingly, participants who had answers to the first test scored higher than control participants, indicating cheating (Cohen's $d \geq 1.16$). In the second part of the experiment, all participants examined a new test for which nobody had the answers, predicted their performance on this test, and completed the test. On this second test, participants who had answers to the first test made higher performance predictions than control participants (Cohen's $d \geq 0.71$). Test performance, however, was equal across groups. Thus, cheating induced overconfidence.

In a later study, Chance, Gino, Norton, and Ariely (2015) investigated the temporal dynamics of *cheating-induced overconfidence*. Students and community members from a paid subject pool ($N = 71$) completed four general-knowledge tests with 10 items. Half the participants received answers to the first test. Again, participants who had answers to a test scored higher on this test than control participants (Cohen's $d \geq 1.58$), indicating cheating. Also, participants who had answers to the first test overpredicted their test performance on the following two tests (Cohen's $d \geq 0.35$). On the fourth test, however, their predictions were not overconfident anymore. Experiment 2 ($N = 148$) showed that when people could cheat again on the third test, overconfidence was fully reinstated on the fourth test (Cohen's $d = 0.23$).

One might wonder whether using answers printed at the bottom of a test indeed constitutes cheating. Chance et al. (2015), however, reported evidence in favor of this interpretation. The researchers recruited $N = 65$ community members via Amazon's Mechanical Turk and provided them with a description of their research paradigm and results. When asked to describe the test takers who had answers to the test, 86% of participants used the words "cheating", "dishonest", "unethical", or synonyms of these words. Also, when rating the extent to which the test takers' behavior constituted cheating on a scale from 1 (definitely not cheating) to 10 (definitely cheating), the modal response was 10 ($M = 6.98$). In contrast, a new sample of $N = 64$ community members who read about participants in the control condition gave a modal rating of 1 ($M = 2.50$). These results indicate that higher scores on tests to which participants have answers constitute cheating as per Chance and colleagues' (2015) experimental instructions.

In sum, previous research suggests that cheating on tests impairs the accuracy of overall performance predictions or *global-level predictions*. However, it is still unclear how profoundly cheating harms monitoring accuracy. In particular, cheating might leave the accuracy of predictions about individual items on a test intact. Unlike global-level predictions, people's *item-level predictions* mainly reflect their experiences with individual items. This often results in participants discounting their metacognitive knowledge and beliefs (Bjork, Dunlosky, & Kornell, 2013; Undorf & Erdfelder, 2015). Item-level predictions have been found repeatedly to be more accurate than global-level predictions. For instance, judgments of text comprehension are more accurate when made for specific item-level terms than for whole global-level passages of text (Dunlosky & Lipko, 2007); moreover, preschoolers' item-level judgments of learning are more accurate than their global-level judgments of learning (Lipowski, Merriman, & Dunlosky, 2013). In other cases, global-level predictions reveal that people have relevant knowledge but fail to use this knowledge when making item-level predictions (Ariel, Hines, & Hertzog, 2014; Hertzog, Price, & Dunlosky, 2008). Given that global-level predictions and item-level predictions are only loosely connected, it is possible that cheating on a test impairs the accuracy of global-level predictions but leaves the absolute and relative accuracy of item-level predictions intact. If so, people who cheat on a test may still be able to accurately monitor their knowledge of individual items, which might reduce the detrimental effects of impaired monitoring. Alternatively, it is possible that cheating impairs the accuracy of both global- and item-level predictions.

The current study aims to determine how profoundly cheating on tests harms monitoring accuracy. To resolve this question, we borrowed the paradigm from Chance et al. (2015, 2011), where cheating is operationally defined as increases in test performance when participants have answers to a test. We used this paradigm across four experiments to systematically investigate whether cheating on a test would impair (1) the absolute accuracy of global-level performance predictions and (2) the absolute and relative accuracy of item-level performance predictions. Experiments 1 and 2 addressed the effects of test cheating on the absolute accuracy of global-level performance predictions. Experiment 3 addressed the effects of cheating on the absolute and relative accuracy of item-level performance predictions. In Experiment 4, we compared the effects of cheating on global- and item-level performance predictions within a single experiment.

2. Experiment 1

In Experiment 1, we sought to replicate the finding that cheating on a test impairs the accuracy of global-level predictions of one's performance on later tests (Chance et al., 2015). As in Chance and colleagues' (2015) study, participants completed four general-knowledge tests. Before completing each test, participants examined it and predicted their score. While Chance et al. (2015) provided half the participants with answers on Test 1 (Experiment 1) or on Tests 1 and 3 (Experiment 2), we provided participants with answers to either the first test (Answers Test 1 group) or the third test (Answers Test 3

group). This allowed us to compare effects of cheating on global-level predictions within and between participants. Based on Chance and colleagues' (2015) results, we had three specific predictions: (1) Participants will cheat on the test with answers; (2) After completing a test with answers, participants will make higher global-level predictions for subsequent tests without answers, resulting in (3) cheating-induced overconfidence in global-level predictions. If so, we should see that predictions exceed scores on the tests following the test with answers. We made no specific predictions about the temporal course of cheating-induced overconfidence. Fig. 1 illustrates our predictions. We pre-registered our predictions and analyses for Experiments 1 to 3 prior to running Experiment 1 (available at https://tinyurl.com/y5llg6lf) and pre-registered our predictions and analyses for Experiment 4 prior to running Experiment 4. For the sake of completeness, we made minor changes to several pre-registered analyses. We mention all deviations from our pre-registered plans below and note any discrepancies in results. Data from all experiments are available at https://tinyurl.com/y55jv9lf.

2.1. Method

2.1.1. Participants

In this experiment and in Experiments 2 and 3, we aimed for $N \geq 82$ to obtain a statistical power of $(1 - \beta) = 0.80$ to detect medium-sized effects ($f = 0.25$) in repeated measures ANOVAs with $\alpha = 0.05$ (all power analyses conducted via G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007). We recruited 149 participants via Amazon's Mechanical Turk and randomly assigned them to the Answers Test 1 and Answers Test 3 groups. We excluded 39 participants who did not complete the experiment, 3 participants who requested that their data be discarded, and 17 participants who indicated that they used a search engine and/or received help from a friend. This left a final sample of $N = 90$, with 42 participants in the Answers Test 1 group and 48 participants in the Answers Test 3 group. The sample was 51.11% female with a mean age of 37.42 ($SD = 11.62$) years. Participants were ethnically Caucasian (72.22%), Asian/Asian-American (11.11%), Hispanic/Latin American (8.89%), African-American (3.33%), Indian (2.22%), Native American (1.11%), and South Asian (1.11%). Most participants were in North America (86.05%), with the remainder residing in Asia (10.47%) and South America (2.33%). Most participants reported that English was their first (87.70%) and their primary language (89.90%).

Participants received $0.50 USD for completing the experiment and learned that they would receive a performance-based bonus of $0.01 per correct answer, though all participants received the full $0.40 bonus.[1]

2.1.2. Materials

We used the same four general-knowledge tests comprised of 10 questions each used by Chance et al. (2015). Questions were of medium difficulty and covered a broad range of topics, including biology, film, geography, and history. This corresponds well with the multi-faceted nature of general knowledge (Irwing, Cammock, & Lynn, 2001; Rolfhus & Ackerman, 1999). Cronbach's α reliabilities ranged from 0.57 to 0.83 ($M = 0.71$, $SD = 0.11$). The order of tests was counterbalanced across participants.

2.1.3. Procedure

All participants completed four general-knowledge tests. All questions from one test appeared simultaneously on the computer screen, with no time limits on making predictions or answering questions. Prior to answering the questions, participants examined the test and predicted their score by typing any whole number from 0 to 10. For the Answers Test 1 group, a box labeled *correct answer* appeared above each question on the screen where participants made the predictions for Test 1 and on the screen where participants answered the respective questions. When participants hovered the mouse over the box, the correct answer appeared. Instructions read, "You can check your answers as you go, but please do your own work." For the Answers Test 3 group, answers appeared the same way on Test 3. Fig. 2 depicts an example of what participants saw on a test with answers. Note that mouse positions could not be recorded because, at the time, there was no readily available way to implement mouse position tracking via Qualtrics survey software. On average, the experiment took 16 min to complete (based on all participants who completed the study, $M = 15.59$, $SD = 7.76$).

2.2. Measures and data analysis

Tests were scored by assigning 1 point for each correct response, resulting in scores between 0 and 10. Answers were considered correct if they were identical to the correct answer (e.g., "Meryl Streep") or unambiguously indicated that participants knew the correct answer (e.g., "Streep", "Ms. Streep", or "Meryl Streepe"). We then converted test scores to percentages. People's performance predictions were also converted to percentages.

If people cheat, scores on tests with answers should exceed those on tests without answers (see Fig. 1). We evaluated this prediction in a mixed ANOVA on test scores with test number (1, 2, 3, 4) as a within-subjects factor and answers test group (1, 3) as a between-subjects factor. In this analysis, cheating should produce a significant interaction. To more closely examine the predicted interaction, we tested separately for each group whether test scores varied across tests, using one-way ANOVAs with test number (1, 2, 3, 4) as a within-subjects factor. We followed up on significant effects of test number using pairwise comparisons (Bonferroni corrected t tests with $p < .008$) and expected to find higher scores on Test 1 than on the other tests in the Answers Test 1 group and higher scores on Test 3 than on the other tests in the Answers Test 3 group.[2]

If cheating on a test results in overconfident global-level predictions for later tests, predicted test scores on the tests after the test with answers should be inflated (see Fig. 1). We evaluated this prediction in a mixed 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) ANOVA on predictions, expecting a significant interaction. Also, cheating-induced overconfidence should produce significant results in separate one-way ANOVAs for each group with test number as a within-subjects factor. Finally, pairwise comparisons (Bonferroni corrected t tests with $p < .008$) should reveal higher predictions on Test 2 than on the other tests in the Answers Test 1 group and higher predictions on Test 4 than on the other tests in the Answers Test 3 group.[3]

Cheating-induced overconfidence should reduce the absolute accuracy of predictions (see Fig. 1). To evaluate this prediction, we conducted a 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) × 2 (Measure: score, prediction) mixed ANOVA. Cheating-induced overconfidence should result in a significant three-way interaction among test

[1] The average hourly wage was $3.46 (based on the experiment's mean duration of 15.59 min, see below). This is above the estimated mean and median hourly wages of workers on Amazon's Mechanical Turk ($3.13 and $1.77, respectively; see Hara et al., 2018) and was approved by our Institutional Review Board. Admittedly, however, it is below minimum wage, at least for participants residing in Northern America.

[2] The pre-registered analysis did not include the omnibus factorial test.

[3] The pre-registered analysis involved separate within-subjects ANOVAs by group that compared only tests without answers.

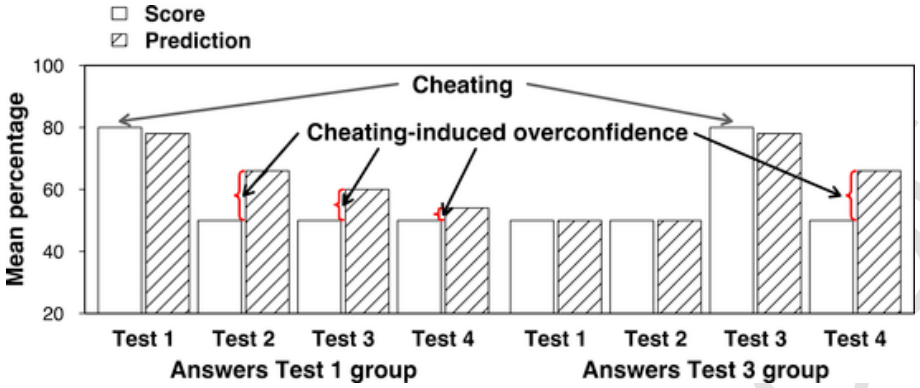


Fig. 1. Hypothetical data illustrating predictions for Experiment 1. The Answers Test 1 group had answers to Test 1 and the Answers Test 3 group had answers to Test 3. Higher scores on tests with answers indicate that participants cheat on those tests. Predictions exceeding scores on the tests following the test with answers indicate cheating-induced overconfidence.

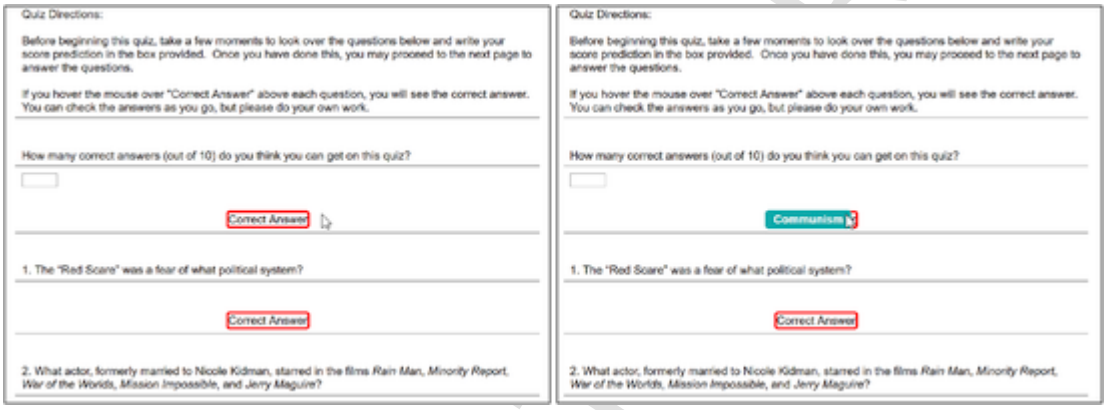


Fig. 2. Example prediction screen for a test with answers. Correct Answer buttons were displayed above each question (left panel). Answers were displayed when participants moved their mouse cursor onto the button (right panel).

number, answers test group, and measure. To more closely examine the predicted interaction, we tested separately for each group whether absolute accuracy varied across tests, using 4 (Test number: 1, 2, 3, 4) × 2 (Measure: score, prediction) repeated-measures ANOVAs. Finally, separate *t* tests were conducted to compare scores and predictions on each test.[4]

To facilitate comparisons of our results to those obtained by Chance et al. (2011), the supplemental materials report comparisons of scores and predictions across answers test groups.

2.3. Results

There were no missing data. Fig. 3 shows actual and predicted test scores for Tests 1 to 4 in the Answers Test 1 and Answers Test 3 groups (see Table S1 in the supplemental materials for descriptive statistics in tables).

2.3.1. Test scores

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on test scores revealed a significant main effect of test number, $F(3, 86) = 26.74$, $p < .001$, $\eta^2_p = .48$, and a significant interaction, $F(3, 86) = 33.31$, $p < .001$, $\eta^2_p = .54$, but no main effect of answers test group, $F(1, 88) = 2.29$, $p = .134$, $\eta^2_p = .03$. Separate analyses for each group revealed that, in the Answers Test 1 group, scores varied across tests, $F(3, 39) = 28.24$, $p < .001$, $\eta^2_p = .69$, with planned comparisons revealing Test 1 > Test 2 = Test 3 = Test 4 (see Table S2 in the supplemental materials for inferential statistics and effect sizes). In the Answers Test 3 group, scores varied across tests, $F(3, 45) = 17.54$, $p < .001$, $\eta^2_p = .54$, with Test 3 > Test 1 = Test 2 = Test 4 (see Table S2 for inferential statistics and effect sizes). Thus, participants from both groups cheated on tests with answers.

2.3.2. Predictions

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on global-level predictions revealed a significant main effect of test number, $F(3, 86) = 10.30$, $p < .001$, $\eta^2_p = .26$, and a significant interaction, $F(3, 86) = 5.96$, $p < .001$, $\eta^2_p = .17$, but no main effect of answers test group, $F(1, 88) = 1.37$, $p = .246$, $\eta^2_p = .02$. Separate analyses for each group revealed that, in the Answers Test 1 group, predictions varied across tests, $F(3, 39) = 8.07$, $p < .001$, $\eta^2_p = .38$, with Test 1 > Test 2 = Test 3 = Test 4 (see Table S2 for inferential statistics and effect sizes). In the Answers Test 3 group, predictions varied across tests, $F(3, 45) = 3.66$, $p = .019$, $\eta^2_p = .20$, with Test 3 > Test 2 = Test 4 and Test 3 = Test 1 (see Table S2 for inferential statistics and effect sizes). Thus, participants made higher predictions on tests with answers than on tests without answers.

2.3.3. Accuracy of predictions

The absolute accuracy of global-level predictions was evaluated in a 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) × 2 (Measure: score, prediction) mixed ANOVA. Significant main effects of test number, $F(3, 86) = 28.94$, $p < .001$, $\eta^2_p = .50$, and measure, $F(1, 88) = 6.92$, $p = .010$, $\eta^2_p = .07$, were qualified by significant interactions between test number and answers test group, $F(3, 86) = 23.82$, $p < .001$, $\eta^2_p = .45$, between test number and measure, $F(3,$

[4] The pre-registered analysis involved separate ANOVAs for each answers test group that included only tests without answers. The reported *t* tests were pre-registered except for the one comparing predictions and scores on the tests with answers.

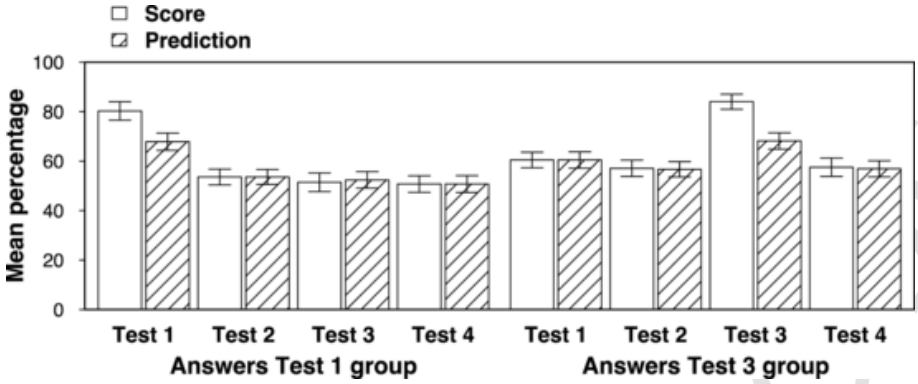


Fig. 3. Mean scores and global-level predictions in Experiment 1. The Answers Test 1 group had answers to Test 1 and the Answers Test 3 group had answers to Test 3. Higher scores on tests with answers indicated cheating. On the tests after the test with answers, predictions were well calibrated (no cheating-induced overconfidence). Error bars represent one standard error of the mean.

86) = 3.75, p = .014, η^2_p = .12, and among test number, answers test group, and measure, $F(3, 86) = 8.15$, $p < .001$, $\eta^2_p = .22$. No other effects were significant, all $F \leq 2.06$, $p \geq .155$. We followed up on the interactions with separate ANOVAs for each group. In the Answers Test 1 group, there was a significant main effect of test number, $F(3, 39) = 24.94$, $p < .001$, $\eta^2_p = .66$, and a significant interaction, $F(3, 39) = 4.23$, $p = .011$, $\eta^2_p = .25$, but no effect of measure, $F(1, 41) = 2.10$, $p = .155$, $\eta^2_p = .05$. Separate t tests revealed significant underconfidence on Test 1, but good calibration on the other tests (see Table S1 for inferential statistics and effect sizes). In the Answers Test 3 group, there were significant main effects of test number, $F(3, 45) = 12.95$, $p < .001$, $\eta^2_p = .46$, and measure, $F(1, 47) = 5.30$, $p = .026$, $\eta^2_p = .10$, and a significant interaction, $F(3, 45) = 5.40$, $p = .003$, $\eta^2_p = .27$. Separate t tests revealed significant underconfidence on Test 3, but good calibration on the other tests (see Table S1 for inferential statistics and effect sizes). Overall, predictions were underconfident on tests with answers but well calibrated on tests without answers. Thus, we did not find cheating-induced overconfidence.

2.4. Discussion

Experiment 1 replicated prior work showing that participants cheated when they had answers to a test; however, we failed to replicate that cheating led to subsequent overconfidence (Chance et al., 2015, 2011). Moreover, although participants made higher global-level predictions on tests with answers, their predictions for these tests were underconfident. We used similar materials and procedures as Chance et al. (2015, 2011). Nevertheless, procedural differences may explain our failure to find cheating-induced overconfidence. Experiment 1 was an online study, whereas Chance et al. (2015, 2011) reported classroom and lab studies. We did not place time limits on the tests (as in Chance et al., 2015, 2011). Unlike Chance et al. (2015, 2011), we did not instruct participants to score their tests. Maybe, self scoring increases the impact of current performance on global-level predictions by highlighting current performance.

It is unlikely that our failure to find cheating-induced overconfidence was due to issues with the quality of data collected from Amazon's Mechanical Turk. In particular, we observed no cases of entire tests left blank or consistent nonsense answers. In contrast, every responder typed in answers to questions from each of the four tests and all incorrect answers were related to the questions (e.g., questions about a famous person were answered with names). This argues against the possibility that responses came from artificial intelligence systems (bots, see Bai, 2018; Dreyfuss, 2018). Also, we identified only four potential repeat responders who had the same IP address and/or geolocation (latitude/longitude) as a previous respondent. Excluding these participants did not change the reported results. Together, these observations suggest that data quality in Experiment 1 was good.

Also, it is unlikely that we did not find cheating-induced overconfidence because participants came from the general population. Given that Experiment 1 participants were older than typical student samples, it is possible that they less often take tests in everyday life. However, if these people cheat on tests, as they did in this study, potentially harmful effects of cheating on monitoring accuracy should nevertheless become evident.

Experiment 2 attempted to find cheating-induced overconfidence in a classroom study with student participants, time limits on tests, and participant self-scoring.

3. Experiment 2

Experiment 2 was identical to Experiment 1 with the exception that participants were students who completed the experiment in five different classrooms at the beginning of class time. Also, all sections of the experiment were timed and participants self-scored each test immediately after answering it. We expected to find that cheating on a test would produce overconfident global-level predictions for subsequent tests, as found in prior work (Chance et al., 2015). Furthermore, we asked participants to remember their score on each test at the end of the experiment to see whether cheating would bias people's memories of their own test performance.

3.1. Method

3.1.1. Participants

We recruited 93 students (Age: M = 20.93, SD = 5.22, 59% female, 38% male, 3% unreported) from five different classrooms at a mid-sized Western Canadian university. Participants were randomly assigned to the Answers Test 1 and Answers Test 3 groups. We excluded the data of six participants (6.45%) who failed to give one or more global-level predictions, leaving 45 participants in the Answers Test 1 group and 43 participants in the Answers Test 3 group. Participants received course credit for participation and were told that they would receive a base credit of 0.5% plus performance-based bonus credit of 0.0125% per correct answer. However, all participants received the full 0.5% of bonus credit.

3.1.2. Materials

Materials were identical to Experiment 1. Cronbach's α reliabilities ranged from 0.75 to 0.84 (M = 0.79, SD = 0.04).

3.1.3. Procedure

The procedure was identical to Experiment 1 with the following exceptions. We used paper-and-pencil tests. For the test with answers, all

answers appeared at the bottom of the test. Participants scored each test after completing it (self scores). At the end of the experiment, they were asked to write down their scores from each quiz (remembered scores). All sections of Experiment 2 were timed. On each test, participants had 1 min to make predictions, 3 min to answer the questions, 1 min to self score, and 1 min at the end of the four tests to remember their scores. Two experimenters distributed testing packages to participants. The experimenters then stayed at the front of the room, timed each section via stopwatches or mobile phones, and announced when participants were to turn the page and proceed to the next section. The experimenters observed participants to ensure that they adhered to time limits and worked on the correct section. However, experimenters did not wander around to observe individual participants. The experiment took about 26 min to complete.

3.2. Measures and data analysis

We obtained the same measures as in Experiment 1. In addition, we obtained participants' self scores and the scores they remembered for each test. Self scores and remembered scored were converted to percentages and then submitted to the same analyses as actual test scores. Data analysis was identical to Experiment 1.

3.3. Results

There were no missing data. Fig. 4 shows actual and predicted test scores for Tests 1 to 4 in the Answers Test 1 and Answers Test 3 groups (see Table S1 in the supplemental materials for descriptive statistics in tables). Self scores and remembered scores appear in the supplemental materials. Both measures were virtually identical to actual test scores and all analyses revealed the same results as those on the actual test scores reported here.

3.3.1. Test scores[5]

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on test scores revealed a significant main effect of test number, $F(3, 84) = 23.89$, $p < .001$, $\eta^2_p = .46$, and a significant interaction, $F(3, 84) = 27.13$, $p < .001$, $\eta^2_p = .49$, but no main effect of answers test group, $F < 1$. Separate analyses for each group revealed that scores varied across tests, Answers Test 1 group: $F(3, 42) = 21.48$, $p < .001$, $\eta^2_p = .61$, with Test 1 > Test 2 = Test 3 = Test 4; Answers Test 3 group: $F(3, 40) = 12.31$, $p < .001$, $\eta^2_p = .48$, with Test 3 > Test 1 = Test 2 = Test 4 (see Table S2 for inferential statistics and effect sizes). Thus, as in Experiment 1, participants cheated on tests with answers.

3.3.2. Predictions[6]

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on global-level predictions revealed a significant main effect of test number, $F(3, 84) = 7.35$, $p < .001$, $\eta^2_p = .21$, and a significant interaction, $F(3, 84) = 8.35$, $p < .001$, $\eta^2_p = .23$, but no main effect of answers test group, $F < 1$. Separate analyses for each group revealed that predictions varied across tests, Answers Test 1 group: $F(3, 42) = 6.96$, $p < .001$, $\eta^2_p = .33$, with Test 1 > Test 2 = Test 3 = Test 4; Answers Test 3 group: $F(3, 40) = 4.46$, $p = .009$, $\eta^2_p = .25$, with Test 3 > Test 1 = Test 2 = Test 4 (see Table S2 for inferential statistics and effect sizes). Thus, participants made higher predictions on tests with answers than on tests without answers.

3.3.3. Accuracy of predictions[7]

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) × 2 (Measure: score, prediction) mixed ANOVA revealed a significant main effect of test number, $F(3, 84) = 18.03$, $p < .001$, $\eta^2_p = .39$, which was qualified by significant interactions between test number and answers test group, $F(3, 84) = 20.59$, $p < .001$, $\eta^2_p = .42$, between test number and measure, $F(3, 84) = 9.87$, $p < .001$, $\eta^2_p = .26$, and among test number, answers test group, and measure, $F(3, 84) = 10.65$, $p < .001$, $\eta^2_p = .28$. No other effects were significant, all $F \leq 2.23$, $p \geq .139$. We followed up on the interactions with separate ANOVAs for each group. In the Answers Test 1 group, there was a significant main effect of test number, $F(3, 42) = 18.76$, $p < .001$, $\eta^2_p = .57$, and a significant interaction, $F(3, 42) = 9.74$, $p < .001$, $\eta^2_p = .41$, but no effect of measure, $F(1, 44) = 1.88$, $p = .177$, $\eta^2_p = .04$. Separate t tests revealed significant underconfidence on Test 1, significant overconfidence on Test 2, and good calibration on Tests 3 and 4 (see Table S1 for inferential statistics and effect sizes). In the Answers Test 3 group, there was a significant main effect of test number, $F(3, 40) = 8.26$, $p < .001$, $\eta^2_p = .38$, and a significant interaction, $F(3, 40) = 5.22$, $p = .004$, $\eta^2_p = .28$, but no effect of measure, $F < 1$. Separate t tests revealed good calibration on Tests 1 and 2, marginal underconfidence on Test 3, and overconfidence on Test 4 (see Table S1 for inferential statistics and effect sizes). Thus, both groups' predictions were underconfident on tests with answers but overconfident on the test immediately after the test with answers, demonstrating cheating-induced overconfidence.

3.4. Discussion

Unlike Experiment 1, this experiment replicated Chance and colleagues' findings (2015, 2011): Cheating on a test produced overconfident global-level predictions for the test that followed. One slight difference to Chance and colleagues' findings (2015, 2011) was that our cheating-induced overconfidence did not persist beyond the test immediately after cheating. Nevertheless, cheating on a test impaired the accuracy of global-level predictions. We found no indication that cheating biased people's memories of their test performance. Thus, remembered scores will not be discussed further.

Given that Experiment 2 revealed cheating-induced overconfidence in global-level predictions, Experiment 3 investigated whether cheating would also harm the accuracy of item-level predictions.

4. Experiment 3

As we mentioned in the introduction, global- and item-level predictions are only loosely connected: Item-level predictions are often more accurate (e.g., Dunlosky & Lipko, 2007; Lipowski et al., 2013). Thus, it is possible that, even when global-level predictions are subject to cheating-induced overconfidence, people are still able to accurately monitor their knowledge of individual items. If so, detrimental effects of impaired monitoring might be greatly reduced. However, it is also possible that cheating impairs the accuracy of item-level predictions. Cheating-induced overconfidence – as was found for global-level predictions – would indicate that cheating impairs the absolute accuracy of item-level predictions. Alternatively or additionally, cheating might harm the relative accuracy of item-level predictions, that is, impair people's ability to distinguish between correct and incorrect responses. As we noted in the Introduction, high absolute and relative accuracy are distinct aspects of monitoring accuracy that foster learning and memory.

[5] The pre-registered analysis did not include the omnibus factorial test.

[6] The pre-registered analysis involved separate within-subjects ANOVAs by answers test group that compared only tests without answers.

[7] The pre-registered analysis involved only tests without answers. It revealed a main effect of measure such that average predictions were higher than average scores.

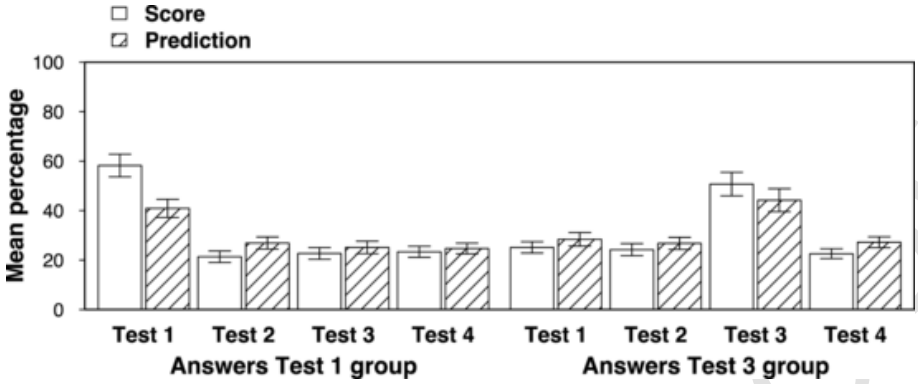


Fig. 4. Mean scores and global-level predictions in Experiment 2. The Answers Test 1 group had answers to Test 1 and the Answers Test 3 group had answers to Test 3. Higher scores on tests with answers indicated cheating. On the test immediately after the test with answers, predictions were overconfident (cheating-induced overconfidence). Error bars represent one standard error of the mean.

In Experiment 3, we investigated the effects of cheating on the absolute and relative accuracy of item-level performance predictions. For each test, participants predicted their chances of answering each individual question correctly. As in the previous experiments, all participants completed four general-knowledge tests and had answers to Test 1 or Test 3. If cheating-induced overconfidence generalizes to item-level predictions, we should find overconfidence on tests that follow a test with answers. Additionally or alternatively, it is possible that cheating impairs people's ability to assess the relative difficulty of questions. If so, we should find lower relative accuracy of item-level predictions for tests that follow on a test with answers.

We will index relative accuracy by a widely used measure: the within-person Goodman-Kruskal gamma correlation between judgments and performance (e.g., Baars, Vink, van Gog, de Bruin, & Paas, 2014; Bröder & Undorf, 2019; Dunlosky & Thiede, 2012). Gamma is a nonparametric correlation coefficient that was identified as superior to other measures of relative accuracy by Nelson (1984). It is widely applicable, makes no scaling assumptions beyond the ordinal level, and can attain maximum values independent of ties and the level of memory performance (see also Gonzalez & Nelson, 1996; for criticism and alternative measures, see; Bröder & Undorf, 2019; Murayama, Sakaki, Yan, & Smith, 2014; Schraw, 2009). Positive gamma correlations, indicating good relative accuracy, result when participants assign higher item-level predictions to correct than to incorrect responses. We expected significantly positive gamma correlations on all tests. If cheating impairs relative accuracy, we should find reduced gamma correlations after the test with answers.

4.1. Method

4.1.1. Participants

We recruited 104 students (Age: $M = 22.58$, $SD = 6.42$, 78% female) at a mid-sized Western Canadian university. Students participated in groups in a lab classroom (1–10 participants per session, $M = 3.29$, $SD = 2.67$), and were randomly assigned to the Answers Test 1 and Answers Test 3 groups. We excluded the data of two participants (1.92%) who failed to make 20 or more item-level predictions, leaving 45 participants in the Answers Test 1 group and 57 participants in the Answers Test 3 group. Participant compensation and performance-based bonuses were the same as in Experiment 2.

4.1.2. Materials

Because participants in Experiment 2 performed worse on the general-knowledge test than participants in prior studies, we selected a new, easier set of 40 general-knowledge questions that included questions used in the previous experiments and new questions from a pool compiled by Tauber, Dunlosky, Rawson, Rhodes, and Sitzman (2013). Based on a pilot study with $N = 83$ students, we selected 40 questions with percentage correct ranging from 32.36% to 58.82%. As in Experiments 1 and 2, questions covered a broad range of topics, including biology, film, geography, history, and sports. We compiled four tests of equal difficulty ($M = 46.27\%$–47.59%, $SD = 8.02$–8.70, *Min*: 32.39%–34.78%, *Max* = 57.75%–58.86%). Cronbach's α reliabilities ranged from 0.65 to 0.85 ($M = 0.77$, $SD = 0.08$).

4.1.3. Procedure

The procedure was identical to Experiment 2 except that we elicited item-level predictions instead of global-level predictions. The prompt "Chance of getting this question correct (0%–100%): ___" appeared beside each question. Participants wrote the probability of responding correctly on each question before answering the questions. Because participants had to make 10 predictions per test, we extended the prediction time to 1.5 min. The experiment took about 28 min to complete. The setting was somewhat different from Experiment 2 in that students participated outside of classes and in smaller groups. Because of this, only one experimenter was present in most sessions.

4.2. Measures and data analysis

Measures were identical to Experiment 2 except that we obtained item-level predictions instead of global-level predictions. Item-level predictions were made on a percentage scale and did not require conversion. We measured the relative accuracy of item-level predictions using within-participants gamma correlations between scores and predictions.

Data analysis was identical to Experiment 1 except that we added two analyses evaluating the relative accuracy of predictions. We first tested whether gamma correlations were significantly positive in all tests and conditions, using one-sample t tests. To test whether cheating impaired the relative accuracy of predictions, gamma correlations were then submitted to a 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA. If cheating impairs the relative accuracy of item-level predictions, we should find a significant interaction between test number and answers test group.

4.3. Results

A total of 12 predicted test scores and 1 remembered test score were missing, resulting in 0.10% missing data. Fig. 5 shows actual and predicted test scores for Tests 1 to 4 in the Answers Test 1 and Answers Test 3 groups (see Table S1 in the supplemental materials for descriptive statistics in tables). As in Experiment 2, self scores and remem-

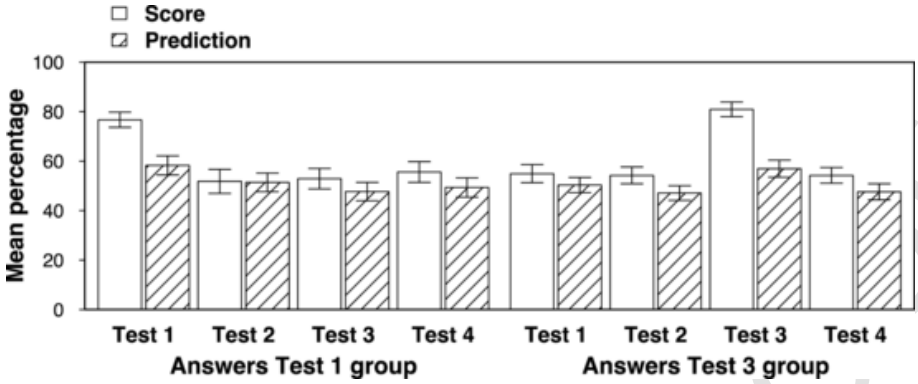


Fig. 5. Mean scores and item-level predictions in Experiment 3. The Answers Test 1 group had answers to Test 1 and the Answers Test 3 group had answers to Test 3. Higher scores on tests with answers than without answers indicated cheating. On the tests after the test with answers, predictions were well calibrated (no cheating-induced overconfidence). Error bars represent one standard error of the mean.

bered scores were virtually identical to actual test scores and revealed the same results as actual test scores (see supplemental materials).

4.3.1. Test scores[8]

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on test scores revealed a significant main effect of test number, $F(3, 98) = 16.01$, $p < .001$, $\eta^2_p = .33$, and a significant interaction, $F(3, 98) = 23.52$, $p < .001$, $\eta^2_p = .42$, but no main effect of answers test group, $F < 1$. Separate analyses for each group revealed that scores varied across tests, Answers Test 1 group: $F(3, 42) = 15.69$, $p < .001$, $\eta^2_p = .53$, with Test 1 > Test 2 = Test 3 = Test 4; Test 3 Answers group: $F(3, 54) = 23.37$, $p < .001$, $\eta^2_p = .57$, with Test 3 > Test 1 = Test 2 = Test 4 (see Table S2 for inferential statistics and effect sizes). Thus, as in the previous experiments, participants cheated on tests with answers.

4.3.2. Predictions[9]

A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA on item-level predictions revealed a marginal main effect of test number, $F(3, 98) = 2.60$, $p = .056$, $\eta^2_p = .07$, and a significant interaction, $F(3, 98) = 5.59$, $p = .001$, $\eta^2_p = .15$, but no main effect of answers test group, $F < 1$. Separate analyses for each group revealed that predictions varied across tests, Answers Test 1 group: $F(3, 42) = 4.16$, $p = .011$, $\eta^2_p = .23$, with Test 1 = Test 2 = Test 4, Test 1 > Test 3, and Test 2 = Test 3 = Test 4; Answers Test 3 group: $F(3, 54) = 4.57$, $p = .006$, $\eta^2_p = .20$, with Test 3 = Test 1, Test 3 > Test 2 = Test 4, and Test 1 = Test 2 = Test 4 (see Table S2 for inferential statistics and effect sizes). Overall, participants made higher item-level predictions on tests with answers than on tests without answers, although differences were smaller than in the previous experiments.

4.3.3. Accuracy of predictions

Absolute accuracy[10] was evaluated in a 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) × 2 (Measure: score, prediction) mixed ANOVA, which revealed significant main effects of test number, $F(3, 98) = 10.45$, $p < .001$, $\eta^2_p = .24$, and measure, $F(1, 100) = 51.32$, $p < .001$, $\eta^2_p = .34$, that were qualified by significant interactions between test number and answers test group, $F(3, 98) = 18.72$, $p < .001$, $\eta^2_p = .36$, between test number and measure, $F(3, 98) = 8.59$, $p < .001$, $\eta^2_p = .21$, and among test number, answers test group, and measure, $F(3, 98) = 12.00$, $p < .001$, $\eta^2_p = .27$. No other effects were significant, all $F \leq 1.43$, $p \geq .234$. We followed up on the interactions with separate ANOVAs for each group. In the Answers Test 1 group, there were significant main effects of test number, $F(3, 42) = 13.67$, $p < .001$, $\eta^2_p = .49$, and measure, $F(1, 44) = 16.98$, $p < .001$, $\eta^2_p = .28$, and a significant interaction, $F(3, 42) = 5.04$, $p = .004$, $\eta^2_p = .27$. Separate *t* tests revealed significant underconfidence on Tests 1 and 4, but good calibration on Tests 2 and 3 (see Table S1 for inferential statistics and effect sizes). In the Answers Test 3 group, there were significant main effects of test number, $F(3, 54) = 17.72$, $p < .001$, $\eta^2_p = .50$, and measure, $F(1, 56) = 37.78$, $p < .001$, $\eta^2_p = .40$, and a significant interaction, $F(3, 54) = 10.58$, $p < .001$, $\eta^2_p = .37$. Separate *t* tests revealed significant underconfidence on Tests 2 to 4, but not on Test 1 (see Table S1 for inferential statistics and effect sizes).

Relative accuracy, indexed by within-participants gamma correlations between scores and predictions, is shown in Table 1. A total of 58 (14.21%) gamma correlations could not be computed because of a lack of variability in scores or predictions. This was most often due to perfect scores on the tests with answers (see Table 1 for the number of participants contributing to each mean correlation).[11] Gamma correlations were positive in all four tests and both groups, indicating good relative accuracy. A 4 (Test number: 1, 2, 3, 4) × 2 (Answers test group: 1, 3) mixed ANOVA[12] revealed no significant effects, all $F < 1$.

Overall, predictions were generally underconfident, with the notable exception of well-calibrated predictions on the two tests following the test with answers in the Answers Test 1 group, and the test immediately after the test with answers in the Answers Test 3 group. Cheating did not affect the relative accuracy of predictions.

4.4. Discussion

In Experiment 3, cheating on a test did not produce overconfidence in item-level predictions. Rather, absolute and relative accuracy of item-level predictions were intact after cheating. These results differ from Experiment 2, in which cheating resulted in overconfident global-level predictions on the subsequent test. Based on these findings, one

[8] The pre-registered analysis did not include the omnibus factorial test.

[9] The pre-registered analysis involved separate within-subjects ANOVAs by answers test group that compared only tests without answers.

[10] The pre-registered analysis did not include the omnibus factorial test and involved only tests without answers. It revealed underconfidence in the Answers test 3 group but not in the Answers test 1 group. These superficial differences do not affect the overall conclusion that cheating did not produce overconfidence.

[11] Because of the relatively large number of ties, we also analyzed relative accuracy in Experiments 3 and 4 using generalized mixed-effects models (Murayama et al., 2014). Results were similar those found with gamma correlations. To remain consistent with the pre-registered analysis and the majority of studies on metacognitive monitoring, we report gamma correlations.

[12] The pre-registered analysis involved separate tests for each group.

Table 1
Means (and standard deviations) of gamma correlations between actual and predicted test scores in experiments 3 and 4.

Experiment and condition	Test number			
	1	2	3	4
Experiment 3				
Answers Test 1 group	.81 (.35), n = 32	.66 (.51), n = 38	.71 (.35), n = 42	.78 (.26), n = 43
Answers Test 3 group	.70 (.44), n = 53	.64 (.46), n = 56	.73 (.50), n = 33	.74 (.37), n = 53
Experiment 4				
Global-level predictions first group			.88 (.26), n = 56	.73 (.49), n = 84
Item-level predictions first group	.80 (.34), n = 64	.74 (.42), n = 107		

Note. All correlations were significant at $p < .001$ in one-sample t tests against 0. n refers to the number of participants contributing to the respective means.

might be tempted to conclude that cheating harms the accuracy of global-level predictions but leaves the accuracy of item-level predictions intact. Experiment 4 directly tested this possibility.

5. Experiment 4

Cheating on a test produced overconfidence in global-level predictions for the subsequent test in Experiment 2 but not in Experiment 1. In Experiment 3, cheating did not produce overconfidence in item-level predictions. Assuming that the lack of cheating-induced overconfidence in Experiment 1 was due to its particular procedure (online study in which participants did not score their tests), a plausible conclusion is that detrimental effects of cheating on monitoring accuracy are limited to global-level predictions. However, caution is warranted, because this conclusion is based on cross-experimental comparisons involving different knowledge tests and procedures. Additionally, Experiment 3 was the first study to investigate the effects of cheating on item-level predictions. Experiment 4 therefore aimed to replicate findings from Experiments 2 and 3 in a single high-powered experiment (pre-registration available at https://tinyurl.com/y3o4u3zn). In this experiment, all participants made global-level predictions on tests with and without answers, and item-level predictions on tests with and without answers. Half the participants made global-level predictions on the first two of four tests (global-level predictions first group), whereas the other half made item-level predictions on the first two of four tests (item-level predictions first group). If harmful effects of cheating on monitoring accuracy are limited to global-level predictions, we should find cheating-induced overconfidence in global-level predictions but intact accuracy of item-level predictions.

5.1. Method

5.1.1. Participants

We aimed for $N \geq 176$ to obtain a statistical power of $(1 - \beta) = 0.95$ to detect medium-sized effects ($d = 0.5$) in independent-samples t tests, and even higher power in all other analyses. We recruited 227 students (Age: $M = 20.6$, $SD = 5.32$, 76% female) at a mid-sized Western Canadian university who were tested in classrooms or groups (1–33 participants per session, $M = 21.23$, $SD = 8.97$). Students were randomly assigned to the global- and item-level predictions first groups. We excluded the data of 13 participants (5.73%) who failed to give all required global-level predictions (8 participants) or more than 10 item-level predictions (5 participants), leaving 97 participants in the global-level predictions first group and 117 participants in the item-level predictions first group. Participant compensation and performance-based bonuses were the same as in Experiments 2 and 3 except for 33 participants who completed the study as part of an on-campus event. These participants were entered into a draw for a $15 coffee shop gift card and gained additional entries into the draw based on their performance.

5.1.2. Materials

Materials were identical to Experiment 3. Cronbach's α reliabilities ranged from 0.75 to 0.81 ($M = 0.77$, $SD = 0.03$).

5.1.3. Procedure

All participants received answers on Tests 1 and 3 and made global-level predictions on two tests (elicited as in Experiment 2) and item-level predictions on the remaining two tests (elicited as in Experiment 3). Participants from the global-first group made global-level predictions on Tests 1 and 2 and item-level predictions on Tests 3 and 4, whereas participants from the item-level-first group made item-level predictions on Tests 1 and 2 and global-level predictions on Tests 3 and 4. In all other respects, the procedure was identical to that of Experiment 3.

5.2. Measures and data analysis

Measures were identical to Experiment 3. To test whether participants cheated, test scores were submitted to a mixed ANOVA with test number (1, 2, 3, 4) as a within-subjects factor and order of predictions group (global-level first, item-level first) as a between-subjects factor. Cheating should produce a significant main effect of test number, with pairwise comparisons (Bonferroni corrected t tests with $p < .008$) revealing higher scores on Tests 1 and 3 than on the other tests in both groups.

If cheating on a test results in overconfident global-level predictions, predicted test scores should be inflated on Test 2 in the global-level first group and on Test 4 in the item-level first group. This should result in a significant interaction between test number (1, 2, 3, 4) and order of predictions group (global-level first, item-level first) in a mixed ANOVA on predictions. Also, it should produce significant results in separate one-way ANOVAs with test number (1, 2, 3, 4) as a within-subjects factor for each group. Finally, pairwise comparisons (Bonferroni corrected t tests with $p < .008$) should reveal higher predictions on Test 2 than on the other tests in the global-level first group and higher predictions on Test 4 than on the other tests in the item-level first group.

To test whether cheating-induced overconfidence reduced the absolute accuracy of global-level predictions, we conducted a 4 (Test number: 1, 2, 3, 4) × 2 (order of predictions group: global-level first, item-level first) × 2 (Measure: score, prediction) mixed ANOVA. Cheating-induced overconfidence in global-level predictions should result in a significant three-way interaction among test number, order of predictions group, and measure. To more closely examine the predicted interaction, we tested separately for each group whether absolute accuracy varied across tests, using 4 (Test number: 1, 2, 3, 4) × 2 (Measure: score, prediction) repeated-measures ANOVAs. Finally, separate t tests were conducted to compare scores and predictions on each test.

We expected that cheating would not affect the relative accuracy of item-level predictions. To test this prediction, we submitted gamma correlations for all tests involving item-level predictions to a 2 (Test: with answers, without answers) × 2 (Order of predictions group: global-level first, item-level first) mixed ANOVA. We expected no significant effects.

5.3. Results

A total of 10 predicted test scores and 1 self score were missing, resulting in 0.05% missing data. Fig. 6 shows actual and predicted test scores for Tests 1 to 4 in the global-first and item-level-first groups (see Table S3 in the supplemental materials for descriptive statistics in tables). As in the previous experiments, self scores and remembered scores were virtually identical to actual test scores and revealed the same results as actual test scores (see supplemental materials).

5.3.1. Test scores

A 4 (Test number: 1, 2, 3, 4) × 2 (Order of predictions group: global-level first, item-level first) mixed ANOVA on test scores revealed a significant main effect of test number, $F(3, 210) = 68.28$, $p < .001$, $\eta^2_p = .49$, but no other significant effects, order of predictions group: $F(1, 212) = 2.65$, $p = .105$, $\eta^2_p = .01$, interaction: $F < 1$. Planned comparisons revealed Test 1 = Test 3 > Test 2 = Test 4 in both groups (see Table S4 in the supplemental materials for inferential statistics and effect sizes). Thus, as in the previous experiments, participants cheated on tests with answers.

5.3.2. Predictions

A 4 (Test number: 1, 2, 3, 4) × 2 (Order of predictions group: global-level first, item-level first) mixed ANOVA on predictions revealed main effects of test number, $F(3, 210) = 26.93$, $p < .001$, $\eta^2_p = .28$, and order of predictions group, $F(1, 212) = 5.52$, $p = .020$, $\eta^2_p = .03$, but no interaction, $F < 1$. Planned comparisons revealed Test 1 = Test 3 > Test 2 = Test 4 in the global-level predictions first group and Test 3 > Test 1 > Test 2, Test 3 > Test 4, and Test 1 = Test 4 in the item-level predictions first group (see Table S4 for inferential statistics and effect sizes). Thus, participants made higher global- and item-level predictions on tests with answers than on tests without answers.

5.3.3. Accuracy of predictions

Absolute accuracy was evaluated in a 4 (Test number: 1, 2, 3, 4) × 2 (Order of predictions group: global-level first, item-level first) × 2 (Measure: score, prediction) mixed ANOVA, which revealed significant main effects of test number, $F(3, 210) = 54.53$, $p < .001$, $\eta^2_p = .44$, order of predictions group, $F(1, 212) = 4.46$, $p = .036$, $\eta^2_p = .02$, and measure, $F(1, 212) = 123.00$, $p < .001$, $\eta^2_p = .37$, and a significant interaction between test number and measure, $F(3, 210) = 24.35$, $p < .001$, $\eta^2_p = .26$. No other interactions were significant, $F \leq 1.83$, $p \geq .177$. Separate t tests revealed significant underconfidence on Tests 1 to 3, but not on Test 4 (see Table S3 for inferential statistics and effect sizes).

Relative accuracy of item-level predictions was again evaluated using gamma correlations. A total of 117 (27.34%) gamma correlations could not be computed. As in Experiment 3, this was most often due to perfect scores on the tests with answers (see Table 1). Gamma correlations were positive in all tests and conditions, indicating good relative accuracy. A 2 (Test: with answers, without answers) × 2 (Order of predictions group: global-level first, item-level first) mixed ANOVA revealed no significant effects, all $F \leq 1.50$, $p \geq .224$.

Overall, global- and item-level predictions were underconfident on all tests, meaning that we did not find cheating-induced overconfidence. Also, cheating did not affect the relative accuracy of item-level predictions, replicating our results in Experiment 3.

6. Discussion

In Experiment 4, cheating on a test harmed neither the accuracy of global-level predictions nor the accuracy of item-level predictions. This indicates that monitoring accuracy is robust against cheating. Intact accuracy of global-level predictions after cheating replicated Experiment 1, but was inconsistent with Experiment 2. Conversely, intact accuracy of item-level predictions replicated Experiment 3 (see Table 2 for an overview of findings).

This finding raises the question of why cheating impaired the accuracy of global-level predictions in Experiment 2 but not in Experiment 4. The observed difference in results seems to be unrelated to the fact that Experiment 4 participants made item-level predictions in addition to global-level predictions, because there was no evidence for cheating-induced overconfidence in participants who made global-level predictions prior to item-level predictions. Also, participants from both experiments came from the same population and all experimental procedures were identical. One difference between Experiments 2 and 4 was that, in Experiment 2, we used the same general-knowledge questions as Chance et al. (2015), while in Experiment 4, we used easier general-knowledge questions that were of similar difficulty for our participants as those used by Chance et al. (2015) for their participants. Our finding of cheating-induced overconfidence in Experiment 2 but not in Experiment 4 may therefore indicate that this effect is closely tied to the specific materials used by Chance et al. (2015). Inconsistent with this possibility, however, Chance et al. (2011, Experiment 1) reported cheating-induced overconfidence on a test of math IQ that used very different questions. We return to this issue in the General Discussion.

Overall, Experiment 4 demonstrated intact accuracy of global- and item-level predictions after cheating on a test, indicating that monitoring accuracy is robust against cheating.

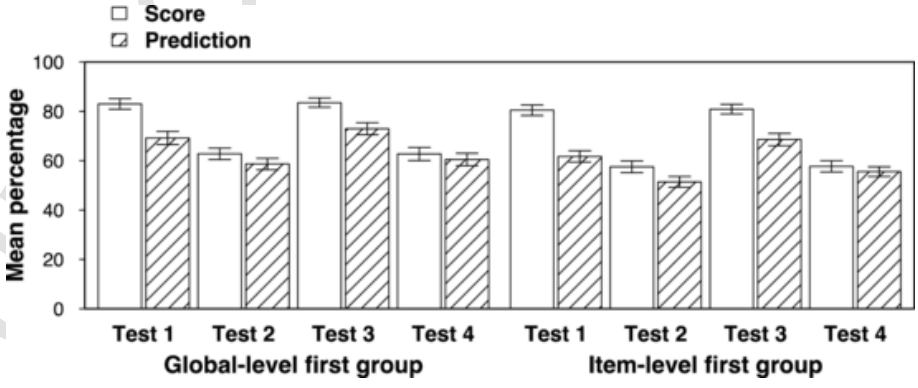


Fig. 6. Mean scores and predictions in Experiment 4. All participants had answers to Tests 1 and 3. The global-level predictions first group made global-level predictions on the first two tests and item-level predictions on the last two tests. The item-level predictions first group made item-level predictions on the first two tests and global-level predictions on the last two tests. Higher scores on tests with answers indicated cheating. On the tests after the tests with answers, predictions were well calibrated (no cheating-induced overconfidence). Error bars represent one standard error of the mean.

Table 2
Summary of experiments 1 to 4.

Experiment	Setting	Materials	Type of predictions	Cheating on tests with answers	Cheating-induced overconfidence	Cheating-impaired relative accuracy
1	Online	Chance et al.'s GK tests	Global-level	Yes	No	
2	Classrooms (during class time)	Chance et al.'s GK tests	Global-level	Yes	Yes	
3	Lab classrooms	Easier GK tests	Item-level	Yes	No	No
4	Classrooms (during class time) and lab classrooms	Easier GK tests	Global-level,	Yes	No	
			Item-level	Yes	No	No

Note. GK = general knowledge.

7. General discussion

The current study addressed whether cheating on a test would harm metacognitive monitoring accuracy. More specifically, we tested whether cheating would (a) produce overconfidence in global-level predictions on future test performance, (b) produce overconfidence in item-level predictions on future test performance, and (c) impair the relative accuracy of item-level predictions. Across four experiments, summarized in Table 2, participants cheated when they had answers to a test, but showed little evidence of cheating-impaired monitoring accuracy. We found cheating-induced overconfidence in global-level predictions in only one of three experiments involving global-level predictions. This indicates that although test cheating can produce overconfident global-level predictions, it is not a consistent source of bias. Neither of the two experiments involving item-level predictions revealed any evidence for harmful effects of cheating on the absolute or relative accuracy of item-level predictions.

We assessed whether our results provided evidence against cheating-induced overconfidence in global- and item-level predictions. We performed Bayesian meta-analyses on the *t* statistics and samples sizes comparing scores and predictions on the test immediately after the test with answers (Rouder & Morey, 2011). We used default specifications in this analysis (Cauchy prior, *r* scale = 0.707). For global-level predictions, the Bayesian meta-analysis indicated that our data were about 33 times more likely under the null than the alternative hypothesis (BF_{10} = 0.03), providing strong evidence against cheating-induced overconfidence in global-level predictions. For item-level predictions, the Bayesian meta-analysis indicated that our data were about 125 times more likely under the null than the alternative hypothesis (BF_{10} = 0.008), providing decisive evidence against cheating-induced overconfidence in item-level predictions. Thus, the current study suggests that monitoring accuracy is robust against cheating. Although cheating may impair learning and memory, our results suggest that any such impairment is not due to impaired monitoring accuracy.

Our results raise the question as to why cheating-induced overconfidence in global-level predictions was less compelling in this study than in the studies by Chance et al. (2015, 2011). Although we cannot answer conclusively, three issues are worth noting. First, the absence of an effect in the current experiments is unlikely due to a lack of statistical power. Our statistical power for detecting medium-sized effects exceeded 0.80 in Experiments 1 to 3 and 0.95 in Experiment 4. Our Experiment 4 is the largest experiment to date assessing effects of cheating on performance predictions for future tests. Second, our finding of cheating-induced overconfidence in Experiment 2 indicates that Chance and colleagues' (2015, 2011) results may be replicable. But the fact that we obtained cheating-induced overconfidence only when running a classroom study that used exactly the same materials as Chance et al. (2015) suggests at least that cheating-induced overconfidence is not as robust or generalizable as one might expect. Finally, one might argue that cheating-induced overconfidence in participants who cheated on tests with answers was counteracted by underconfidence in participants who did not cheat. Although we cannot entirely rule out this possibility, an exploratory analysis argued against it. In this analysis, we scored participants as cheaters if their scores on tests with answers were higher than their maximum score on tests without answers. By this definition, between 60.19% (Experiment 3) and 83.18% (Experiment 4) of participants qualified as cheaters. Importantly, there was no sign of cheating-induced overconfidence in the cheaters' predictions, all $t \leq 2.03$, all $p \geq .050$.

To improve the generalizability of our results, we recruited participants from two different populations[13] and used two different sets of general-knowledge test materials. Nevertheless, using general-knowledge questions across all experiments may have limited the generalizability of our findings in certain respects. In particular, we cannot exclude the possibility that cheating may harm monitoring accuracy for materials other than general-knowledge questions. One might speculate that detrimental effects of cheating on monitoring accuracy may be found when tests are very similar. Although each of our general-knowledge tests covered similar areas of general knowledge, participants may have still felt that similarity between tests was limited (e.g., the two geography questions "What is the capital of and largest city in Japan?" and "In what U.S. state is Atlantic City located?" may appear very different to a person who is more knowledgeable about American geography than about Asian geography). If so, cheating may harm monitoring accuracy when the similarity between tests is higher, such as for tests on course-based materials. Similarly, it is an open question whether cheating on tests of newly learned materials may harm monitoring accuracy. Maybe it is easier to distinguish whether accurate responses resulted from cheating or from knowledge stored in one's semantic mem-

[13] Participants from the two populations differed with respect to both age and location. We do not know, however, to what extent they differed in education or socioeconomic status.

ory than to distinguish whether accurate responses resulted from cheating or from recent studying. If so, cheating may harm monitoring accuracy for future tests on recently studied information. Finally, increasing the spacing between tests in time might decrease people's memory of the reasons for their good performance and, consequently, induce harmful effects of cheating on monitoring accuracy.

Importantly, however, apart from studies demonstrating that people expect their future level of performance to match their current level of performance (e.g., Critcher & Rosenzweig, 2014; Kornell & Hausman, 2017) and a single experiment by Chance et al. (2011, Experiment 1), there currently is no research suggesting that cheating on tests with materials other than general-knowledge questions may impair monitoring accuracy.

Although these speculations are intriguing and may merit future research, using tests with very similar or newly learned materials may create an issue in the interpretation of overconfidence after test cheating. Specifically, whenever answering one test question strengthens memory for answers to other test questions, overconfidence may occur because people overestimate the benefits of looking up answers at the bottom of the page and, at the same time, underestimate the benefits of trying to recall answers from memory. Consistent with this possibility, learners often regard restudying as more effective than retrieval practice, even though actual performance shows the opposite pattern (Karpicke, 2009; Kornell & Son, 2009). We think that underestimating the benefits of retrieval practice may not explain cheating-induced overconfidence when using general-knowledge tests where retrieval practice is not critical. However, it might be essential to consider neglecting the benefits of retrieval practice as an alternative explanation for cheating-induced overconfidence when using test materials other than general-knowledge questions.

Our finding of overconfident global-level predictions after cheating in Experiment 2 but not in Experiments 1 or 4 might be taken to suggest that difficult items promote cheating-induced overconfidence (mean scores across tests without answers were 22.85% in Experiment 2 but 55.33% in Experiment 1 and 60.00% in Experiment 4). Note, however, that this explanation is not fully consistent with Chance et al. (2015) finding of cheating-induced overconfidence with scores around 50%. Clearly, further research is needed to confirm a potential link between test difficulty and harmful effects of cheating on monitoring accuracy.

The current results have obvious educational relevance. Research on academic integrity has identified several ways to reduce but not eliminate test cheating (Bretrag, 2016; McCabe et al., 2012). For instance, McCabe and Trevino (1993) found that academic honor codes reduced the percentage of students who admitted to copying from others on tests from 32% to 13%. In view of the pervasiveness of test cheating, the current results might be comforting. In particular, they argue against the idea that test cheating impairs monitoring accuracy and, consequently, learning and memory. Considering that test cheating has other adverse consequences, however, this does not condone cheating.

Of course, there are several potential challenges and limitations to the generalizability of the present results to real-world cheating. First, in our experiments, cheating may have been more ambiguous and easier than in many real-world tests, mainly because we provided answers at the bottom of the test. Consequently, our design may have inflated cheating as compared to real-world tests. At the same time, there is little doubt that using provided answers when working on the test constituted cheating. Doing so clearly violated test instructions and is regarded as cheating by community members (Chance et al., 2015). Moreover, gains from cheating in our experiments were presumably smaller than in many real-world tests. Although we made sure to incentivize good performance by a performance-based bonus pay, people may regard good performance on high- and low-stakes real-world tests as more important. Finally, in most real-world tests, test takers are free to choose between different forms of cheating (e.g., copying from another student, use of cheat sheets, searching the Internet). In our studies, we restricted cheating to using the answers provided at the bottom of the test. We supervised participants in the three experiments occurring in classrooms and labs (Experiments 2 to 4). In the online experiment, we excluded all participants who indicated that they violated instructions by using a search engine or receiving help from a friend (Experiment 1, 17 of 109 participants). It is thus possible that we excluded the boldest cheaters from our final sample in Experiment 1. Also, our strict definition of test cheating may have reduced the ecological validity of our study.

A limitation to the current study is that we did not measure cheating at the level of individual participants or at the level of individual test questions. Although higher scores on tests with answers clearly indicate cheating on the group level, they do not tell us whether individual participants cheated on individual test questions. Supplementing the current design with eye movements or, in computerized experiments, mouse clicks would provide good measures of cheating at the individual level. Unraveling cheating at the individual level may be an exciting and worthwhile endeavor for future research. Also, obtaining qualitative measures of participants' approaches to and thoughts about the tests may provide important insights into the reasons and boundary conditions of intact monitoring accuracy after cheating.

In summary, we found no evidence that test cheating harmed learning and memory processes due to impaired monitoring accuracy. Instead, our results demonstrate that monitoring accuracy is robust against potential biasing effects of cheating on a test.

Funding

This work was supported by a Margarete von Wrangell fellowship from the state of Baden-Württemberg to Monika Undorf and by grants from the Canada Research Chairs Program (950-232078) and the Social Sciences and Humanities Research Council of Canada (435-2015-0721) to Daniel M. Bernstein.

Declaration of competing interest

None.

Acknowledgments

We thank Zoë Chance for providing the materials used in Experiments 1 and 2. We thank the instructors who allowed us to test in their classes and Andrew Heubert for his help with data collection in Experiment 2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2019.101295.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. Journal of Personality and Social Psychology, 49, 1621–1630. doi:10.1037/0022-3514.49.6.1621.

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. Journal of Memory and Language, 75, 181–198. doi:10.1016/j.jml.2014.06.003.

Arnold, I. J. M. (2016). Cheating at online formative tests: Does it pay off? Internet and Higher Education, 29, 98–106. doi:10.1016/j.iheduc.2016.02.001.

Baars, M., Vink, S., van Gog, T., de Bruin, A. B. H., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. Learning and Instruction, 33, 92–107. doi:10.1016/j.learninstruc.2014.04.004.

Bai, H. (2018). Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. Retrieved from https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444. doi:10.1146/annurev-psych-113011-143823.

Bretrag, T. (2016). Handbook of academic integrity. Singapore: Springer. doi:10.1007/978-981-287-098-8.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. Acta Psychologica, 197, 153–165. doi:10.1016/j.actpsy.2019.04.011.

Chance, Z., Gino, F., Norton, M. I., & Ariely, D. (2015). The slow decay and quick revival of self-deception. Frontiers in Psychology, 6, 1–6. doi:10.3389/fpsyg.2015.01075.

Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. Proceedings of the National Academy of Sciences, 108, 15655–15659. doi:10.1073/pnas.1010658108.

Critcher, C. R., & Rosenzweig, E. L. (2014). The performance heuristic: A misguided reliance on past success when predicting prospects for improvement. Journal of Experimental Psychology: General, 143, 480–485. doi:10.1037/a0034129.

Dreyfuss, E. (2018). A bot panic hits Amazon's Mechanical Turk. Retrieved from https://www.wired.com/story/amazon-mechanical-turk-bot-panic/.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. Current Directions in Psychological Science, 16, 228–232. doi:10.1111/j.1467-8721.2007.00509.x.

Dunlosky, J., & Metcalfe, J. (2009). Metacognition. Los Angeles, CA: Sage.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. Learning and Instruction, 22, 271–280. doi:10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., & Thiede, K. W. (2012). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. Learning and Instruction, 2–5. doi:10.1016/j.learninstruc.2012.05.002.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175–191. doi:10.3758/BF03193146.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. Psychological Bulletin, 119, 159–165. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8559859.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. Proceedings of the 2018 conference on human factors in computing systems. doi:10.1145/3173574.3174023.

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness? Learning and Individual Differences, 18, 430–445. doi:10.1016/j.lindif.2007.12.002.

Irwing, P., Cammock, T., & Lynn, R. (2001). Some evidence for the existence of a general factor of semantic memory and its components. Personality and Individual Differences, 30, 857–871. doi:10.1016/S0191-8869(00)00078-7.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. Journal of Experimental Psychology: General, 138, 469–486. doi:10.1037/a0017341.

Koriat, A. (2007). Metacognition and consciousness. In Zelazo, P. D., Moscovitch, M., & Thompson, E. (Eds.), Cambridge handbook of consciousness (pp. 289–325). New York, NY: Cambridge University Press.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. Journal of Experimental Psychology: General, 138, 449–468. doi:10.1037/a0017350.

Kornell, N., & Hausman, H. (2017). Performance bias: Why judgments of learning are not affected by learning. Memory & Cognition, 45, 1270–1280. doi:10.3758/s13421-017-0740-1.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. Memory, 17, 493–501. doi:10.1080/09658210902832915.

Lipowski, S. L., Merriman, W. E., & Dunlosky, J. (2013). Preschoolers can make highly accurate judgments of learning. Developmental Psychology, 49, 1505–1516. doi:10.1037/a0030614.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37, 2098–2109. doi:10.1037/0022-3514.37.11.2098.

McCabe, D. L., Butterfield, K. D., & Trevino, L. K. (2012). Cheating in college: Why students do it and what educators can do about it. Baltimore, MD: Johns Hopkins University Press.

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty. The Journal of Higher Education, 64, 522–538. doi:10.1080/00221546.1993.11778446.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. European Journal of Cognitive Psychology, 19(4–5), 494–513. doi:10.1080/09541440701326154.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 1287–1306. doi:10.1037/a0036914.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95, 109–133.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. Metacognition and Learning, 1, 159–179. doi:10.1007/s10409-006-9595-6.

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. Metacognition and Learning, 13, 179–212. doi:10.1007/s11409-018-9183-8.

Passow, H. J., Mayhew, M. J., Finelli, C. J., Harding, T. S., & Carpenter, D. D. (2006). Factors influencing engineering students' decisions to cheat by type of assessment. Research in Higher Education, 47, 643–684. doi:10.1007/s11162-006-9010-y.

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. Psychological Bulletin, 102, 122–138. doi:10.1037/0033-2909.102.1.122.

Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. Learning and Individual Differences, 29, 141–149. doi:10.1016/j.lindif.2012.12.003.

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence, and related traits. Journal of Educational Psychology, 91, 511–526. doi:10.1037/0022-0663.91.3.511.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. Psychonomic Bulletin & Review, 18, 682–689. doi:10.3758/s13423-011-0088-7.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, 4, 33–45. doi:10.1007/s11409-008-9031-3.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. Behavior Research Methods, 45, 1115–1143. doi:10.3758/s13428-012-0307-9.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, 103, 193–210. doi:10.1037/0033-2909.103.2.193.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. Psychonomic Bulletin & Review, 6, 662–667. doi:10.3758/BF03212976.

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. Memory & Cognition, 43, 647–658. doi:10.3758/s13421-014-0479-x.