Metacognition and Learning
https://doi.org/10.1007/s11409-022-09295-x

Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy

Nick Naujoks[1] · **Bettina Harder**[1] · **Marion Händel**[1]

Received: 18 June 2021 / Accepted: 26 February 2022

Abstract
Two studies investigated the testing effects on performance and on metacognitive judgment accuracy in authentic learning settings. Across two educational psychology courses, undergraduate students had the opportunity to voluntarily participate in four different practice tests during the term—without feedback in Study 1 ($N = 201$ students) or with individual corrective feedback in Study 2 ($N = 111$ students). Across studies in real classroom settings with and without feedback, regression analyses indicated that a higher number of taken practice tests were related to higher performance in the final course exam and to two scores of metacognitive judgment accuracy (absolute accuracy and sensitivity). However, students' preparation and post-processing practice tests, their perceived usefulness of tests for monitoring one's performance, and metacognitive specificity differed depending on whether students received feedback or not. Overall, the studies convey considerable evidence on how participation in practice tests is related not only to performance but also to monitoring accuracy in authentic learning settings.

Keywords Testing · Performance · Metacognitive judgments · Feedback

In higher education, final course exams usually take place at the end of the study term and students have to study for several exams taking place in a specific period. Ideally, students study lecture content during the term continuously (Susser & McCabe, 2012) in order to understand and master the high amount of learning content. Such a study behavior would allow students to avoid cramming for the exam or overload in the exam period (Hartwig & Dunlosky, 2012; Susser & McCabe, 2012). To support students in their self-regulated learning process during the term, educators can implement practice tests into regular courses (Ariel & Karpicke, 2018; Cogliano et al., 2019). From a theoretical perspective, practice tests function as a (repeated) retrieval opportunity, which has the potential to

✉ Nick Naujoks
nick.naujoks@fau.de

[1] Department of Psychology, Friedrich-Alexander University Erlangen-Nuremberg, Regensburger Straße 160, 90478 Nuremberg, Germany

empower future learning, recall, and recognition (Adesope et al., 2017; Greving & Richter, 2018; McDaniel et al., 2007; Schwieren et al., 2017). In addition, it is suggested that practice test participation also establishes potentials for learning processes like metacognitive judgment accuracy (Karpicke & Roediger, 2008). Hence, practice tests during term seem a suitable and beneficial learning tool to bring together claims of a continuous self-regulated learning practice and students' request for original exam items (Lyle & Crawford, 2011).

However, it is unclear how to implement practice tests in real classroom settings, especially with regard to the provision of feedback on practice tests (Agarwal et al., 2008; Barenberg & Dutke, 2018) and its impact on metacognitive judgment accuracy (henceforth judgment accuracy) as an indicator for metacognitive monitoring (Barenberg & Dutke, 2021; Cogliano et al., 2019). According to Schraw (2009, p. 33) metacognitive judgment accuracy is "the precision of a judgment about a specific task". On the one hand, tests themselves provide students with feedback on their learning progress through successful or unsuccessful retrieval. On the other hand, even students' who engage into practice tests show inaccurate judgment accuracy, which leads to the conclusion that feedback is essential to strengthen the adequate use of tests regarding judgment accuracy. Therefore, we used practice tests as retrieval practice in two classroom-studies with and without individual corrective feedback and studied its effects on final exam performance and judgment accuracy.

The testing effect

Research results since Abbott's (1909) initial study on recall suggest that pure participation in practice tests without feedback is a powerful learning strategy to improve performance. The testing effect describes the benefit of repeated retrieval in practice tests on performance in a criterial test (Roediger & Karpicke, 2006). In the following, we focus on three potential effects of participation in practice tests on performance and judgment accuracy. First, the *direct testing effect* means that repeated retrieval of identical items and higher performance in a criterial test with the exact same information are related. It has been replicated with various samples, test formats, and design conditions (for a review, see Dunlosky et al., 2013). The majority of (laboratory) studies showed that students who took practice tests performed better on a subsequent retention test with identical items than students who used more common study strategies like restudying (e.g., Kubik et al., 2018; Lehman et al., 2014; Roediger & Karpicke, 2006). Second, this effect transfers to tests with non-identical items that are common in real classroom settings (e.g., Batsell et al., 2016; McDaniel et al., 2012; Moreira et al., 2019; Roediger et al., 2011; Schwieren et al., 2017). Third, current research investigates an *indirect testing effect*, namely that on judgment accuracy (Händel et al., 2020; Tullis et al., 2013). The present study examines testing effects on performance and judgment accuracy. A special focus of the studies is on the role of feedback, which is discussed as a relevant design element of participation in practice tests in real classroom settings (Greving & Richter, 2018; Hattie & Timperley, 2007) and a major moderator of the testing effect on performance (Rowland, 2014; Schwieren et al., 2017).

The testing effect and judgment accuracy

Besides the effects of testing on performance, participation in practice tests is supposed to be associated with several aspects of (self-regulated) learning (Ariel & Karpicke, 2018; Fernandez & Jamet, 2016; Roediger et al., 2011). In particular, practice tests contain metacognitive cues that can help learners to monitor their current level of knowledge or detect knowledge gaps (Barenberg & Dutke, 2018; Fernandez & Jamet, 2016; Kelemen, 2000). Monitoring as a part of the procedural component of metacognition involves a process in which the learner observes and reflects his or her own cognitive processes and evaluates personal progress (Flavell, 1979). Indicators of metacognitive monitoring are so-called metacognitive judgments (henceforth judgments) by means of which students convey whether or not they know what they know (Koriat, 2019). Specifically, judgments of performance can be distinguished into prospective, concurrent, and retrospective judgments as well as into global (referring to a whole test) and item-specific judgments (Dunlosky & Metcalfe, 2009; Hacker et al., 2008). In the case of this study, students provided a dichotomous judgment on whether they think they answered correctly or not after answering each item (retrospective, item-specific judgment). Judgments inherit an evaluative standard and allow researchers to compute several accuracy or calibration scores (Bol & Hacker, 2012). Absolute measures like absolute accuracy (Schraw, 2009) assess students' overall ability to estimate their performance. They represent the precision or resolution (Vuorre & Metcalfe, 2021) of judgments compared to performance. Absolute accuracy indicates the fit of performance p_i and judgment j_i. Values close to zero point to inaccurate monitoring and values close to 100 indicate accurate judgments:

$$\text{Absoluteaccuracy} = 100 - \frac{1}{n}\sum_{i=1}^{n}\left|j_i - p_i\right| \tag{1}$$

Relative accuracy scores available for item-specific judgments allow the assessment of, for example, efficiency, discrimination or agreement and measure the relationship of judgments and performance scores (Schraw et al., 2013). Dichotomous judgments allow the coding according to a 2×2 matrix of item solution and judgment: hit, false alarm, miss, and correct rejection as according to signal detection theory (Green & Swets, 1966). This allowed for calculating sensitivity and specificity (Schraw et al., 2013). Sensitivity indicates the relative frequency of accurately detected correct answers:

$$\text{Sensitivity} = \frac{\sum hits}{(\sum hits + \sum misses)} \tag{2}$$

Specificity indicates the relative frequency of accurately detected incorrect answers:

$$\text{Specificity} = \frac{\sum correctrejections}{(\sum falsealarms + \sum correctrejections)} \tag{3}$$

Research investigating relationships of participation in practice tests and judgment accuracy in laboratory settings or with materials of limited meaningfulness (e.g., associates or non-course related materials) showed that students judged their performance more accurate after repeated test taking (e.g., Ariel & Dunlosky, 2011; Chen et al., 2019; Finn & Metcalfe, 2007; Jönsson et al., 2012; Roediger & Karpicke, 2006; Tullis et al., 2013). Focusing on real classroom settings with higher education students, however, research findings are heterogeneous. Studies examining judgment accuracy on a global level (Bol et al., 2005;

Foster et al., 2017; Hacker et al., 2008) and the level of conceptual knowledge (percentage of correct answers regarding a specific concept; Rivers et al., 2019) did not show improvement in students' monitoring accuracy after experiencing multiple practice tests in their regular courses. In contrast, studies investigating judgment accuracy on an item-specific level and with course-relevant material (Cogliano et al., 2019; Händel et al., 2020) but also with judgments on a global level (Fernandez & Jamet, 2016) found promising evidence on increased accuracy when students engage in test taking. Cogliano et al's (2019) and Händel et al's (2020) research findings suggest that item-specific judgments, in particular, can provide evidence of whether the metacognitive potentials of the practice tests promote recalibration of student judgments. Students should be able to provide an accurate overview of their judgment accuracy based on the assessment of individual items and use this overview to correct any misconceptions. However, this task is difficult for students and they tend to be overconfident regarding their metacognitive judgments accuracy (Cogliano et al., 2019; Händel et al., 2020). A possible assumption is that feedback (either self-generated or externally given by, for example, teachers in real classroom settings) moderates the effect of repeated retrieval on judgment accuracy, as discussed below (Agarwal et al., 2008; Miller & Geraci, 2011).

Practice tests with and without feedback

Feedback contains information for learners to monitor or regulate their learning in future situations. It can trigger confirmation, addition, overwriting, tuning or reconstruction of information in memory (Butler & Winne, 1995) and thus should amplify the direct testing effect by providing opportunities for deeper processing and disclose errors (Butler & Roediger, 2008; Schwieren et al., 2017). However, Adesope et al. (2017) found no significant differences between testing effects with or without feedback in their meta-analysis. It is possible that the effectiveness of the feedback is related to students' perception of the testing situation as meaningful and worth engaging in (Jönsson & Panadero, 2018; Kornell, 2014). Hence, differences in the meaningfulness of settings (e.g., laboratory setting with no relations to personal performance in contrast to a mandatory course setting where students prepare and post-process test content) might be an explanation for the ambiguous role of feedback on practice tests in the testing literature (Adesope et al., 2017; Butler & Roediger, 2008; Rowland, 2014). Summarizing existing studies in this area, such a setting in the context of a lecture is characterized by the following features (Barenberg & Dutke, 2021; Cogliano et al., 2019; Enders et al., 2021; Fernandez & Jamet, 2016; Händel et al., 2020; McDaniel et al., 2007, 2012; Moreira et al., 2019; Raaijmakers et al., 2019):

- voluntary participation in practice tests
- content of practice tests reflects the material of the respective lecture/exam
- practice tests of same format as the criterial test (multiple-choice)
- corrective feedback (correctness of the answer and the correct solution)

In addition, students should continue to process the content of the tests and not merely be interested in gaining insight into the difficulty and structure of the exam (Dutke et al., 2010). However, feedback is not only linked to performance, but can also be beneficial at the level of self-regulation or metacognition, which will be discussed in more detail below (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007).

Feedback and metacognitive judgment accuracy

Regarding judgments, feedback might have different functions and relations compared to test performance. While Greving and Richter (2018), for example, concluded that feedback might disturb the testing effect because of additional exposure to learning materials, Fernandez and Jamet (2016) emphasize its importance for stimulating recalibration of judgment accuracy (at least for true/false feedback when item solution is not given). Engaging in practice tests might be considered as (self-generated) feedback (Fernandez & Jamet, 2016; Hattie & Timperley, 2007). That is, while completing a practice test, students spontaneously generate monitoring judgments functioning as internal feedback that helps them to engage in future learning for the criterion test. Supporting this assumption, studies showed that participation in practice tests leads to more elaborated knowledge, which, in turn, should facilitate judgments (e.g., Endres & Renkl, 2015).

A study reinforcing the assumption of recalibration of judgment accuracy used feedback that only offered information on item correctness (true/false) but did not correct wrong answers (Fernandez & Jamet, 2016). Students performed better when they engaged into practice tests and showed effective self-regulated learning behavior. However, the authors used open answer questions, and, thus, it is not clear whether this transfers to multiple-choice questions (actually, research on testing effects on performance indicates differences due to question format; Butler & Roediger, 2007; McDaniel et al., 2007; McDermott et al., 2014). For example, when solving multiple-choice questions with several distractors, students might be distracted by the alternate answer options when no corrective feedback is provided (Greving & Richter, 2018). Hence, when students are unsure about which answer is the correct one, individual corrective feedback should strengthen the testing effect on judgment accuracy.

Summarizing the challenges of current research on the testing effect, we argue that the effects of repeated retrieval on performance and monitoring need a closer look. This concerns the role of feedback and particularly the investigation of different scores of judgments. On the one hand, practice tests with feedback should help students to monitor their performance in future learning situations as feedback provides them with knowledge about their strengths and weaknesses. On the other hand, students' metacognitive judgment accuracy might benefit from pure participation in practice tests, that is practice tests without feedback. Without having access to corrective feedback, students might generate internal feedback based on the cues during task performance (Koriat et al., 2008) or might regulate their learning and invest more effort.

Research questions and hypotheses

Based on previous research, we aimed to investigate how the self-regulated use of practice tests in an authentic course setting is related to performance and judgment accuracy.

First, we supposed that the testing effect on performance can be detected for pure participation in practice tests and testing with feedback when testing for course-related content.

H1: The more frequently students engage in pure participation in practice tests during the semester the better they perform in a final exam.

H2: The more frequent students engage in practice tests with additional feedback the better they perform in a final exam.

Next, we supposed that students' judgment accuracy and their test-taking behavior regarding course-related practice tests were interrelated.

H3: The more frequent students engage in practice tests the more accurately they provide their judgments in a final exam.
In detail, we suggested that students show higher absolute accuracy and higher efficiency (higher sensitivity and specificity). However, we examined whether the effects on judgment accuracy occur only for practice tests with or without feedback (or both) as an open research question.
Q1: Do the effects of repeated participation in practice tests on judgment accuracy vary for practice tests with or without feedback?
To enlighten the effects further, we studied the preparation, post-processing and perceived usefulness of practice tests with and without feedback as an exploratory research question.
Q2: How is feedback related to preparation and post-processing of retrieval tests?
Q3: Does students' perceived usefulness of participation in practice tests differ between tests with and without feedback?

General method

We conducted two studies to investigate the relationships between participation in practice tests on both exam performance and judgment accuracy. The studies implemented a pretest and a posttest (final exam) as well as four practice test opportunities during the term (via online practice tests) in educational psychology lectures. Every student could decide on his or her own whether to use the self-testing opportunity. That is, we offered every student in the course the possibility to make use of the whole learning and testing material. This was decided due to ethical reasons in a real classroom setting. Hence, we refrained from disadvantaging any student by withholding learning or testing material from them that might be relevant and helpful for the upcoming exam. Study 1 focused on the effects of pure participation in practice tests; Study 2 was concerned with the effects of practice tests plus feedback.

Study 1

Study 1 aimed to examine whether students who engaged more frequently into pure participation in practice tests showed significantly higher test performance (H1) and more accurate judgments of learning (H3).

Method

Study Design

We implemented the study in a regular course setting that lasted for one study term (14 weeks). In the first week of term, students participated in a pretest to assess their

item-specific judgments regarding course-relevant prior knowledge. In addition, we assessed students' gender, grade point average (GPA) and study year.

After each thematic block of the lecture, that is, approximately every three weeks, students had the opportunity to participate in an online practice test referring to the topic of the last weeks. In total, we implemented four practice tests on a university learning platform. Each practice test was available online for ten days. Students could save the items as a PDF file after completing each practice test and did not receive any additional information on their test score.

One week before the course exam took place, that is, after students had the opportunity to participate in all practice tests, we asked them to fill in a questionnaire regarding their preparation and post-processing of the practice tests. Finally, in the last week of term, students participated in the final course exam and provided item-specific judgments. In contrast to the four practice tests, the pretest, the posttest, and the questionnaire were paper and pencil instruments.

Sample

Based on medium effect sizes reported in previous studies, we conducted a power analysis. To detect medium effects of $f^2 = 0.15$ (presuming $\alpha = 0.05$, $1—\beta = 0.95$) in a regression analysis design with two predictors (i.e., pretest variable and frequency of practice tests), a minimum sample size of 107 participants was indicated.

The participants were students of a psychology lecture for undergraduate students. Of the 328 students who took the corresponding exam in educational psychology, $n = 201$ took part in the pretest and provided judgments in the pre- and posttest. Students who provided item-specific judgments regarding their exam had comparable exam performance scores as students who did not voluntarily provide the judgments ($t(326) = 0.89$, $p = 0.37$), that is, the sample seems unbiased in this regard. Most of the participants were first-year students (84.5%). The majority of students were female (78.0%), which is typical for university introductory courses in this field of study. Students' GPA can be regarded as average, $M = 2.50$, $SD = 0.51$, on a scale ranging from 1 to 6 with lower values indicating better grades.

Participation in the practice tests was as follows: 54 students participated in none of the practice tests, 55 students filled in one practice test, 35 students participated in two practice tests, 14 students participated in three practice tests, and 43 students took part in all four practice tests. Participation in practice tests was treated as a continuous variable ranging from 0 (no practice test at all) to 4 (participation in all practice tests). Consistent with empirical evidence on cramming (e.g., Blasiman et al., 2017), students who took one or two practice tests primarily took the last test opportunity before the exam.

Instruments

The study implemented several performance tests (pretest, practice tests, final course exam) and according judgments (pretest, final course exam). All performance measures used the same item format: multiple-choice items with one correct answer out of four answer possibilities.

Prior knowledge (pretest). The prior knowledge test consisted of 12 relatively general knowledge items. It related to content from the just commencing psychology course and was non-identical to the exam items. Students had 15 min to solve the pretest.

Practice tests. The practice tests met all criteria of meaningful learning settings. During the term, students could voluntarily complete four different practice tests, each consisting of 12 items. All tests were online multiple-choice tests and were curricularly valid. The practice tests were based on the course content of the respective previous weeks. The topics of the four practice tests were: 1) empirical methods, 2) educational diagnostics, 3) learning theories, and 4) memory.

Test performance. The final course exam served as an indicator of performance. The implemented exam covered all lecture topics. However, all exam items were unknown to the students, that is, none of the items of the practice tests had been used for the posttest. Students had to complete the exam within 45 min. It consisted of 32 multiple-choice questions (Cronbach's $\alpha = 0.69$).

Judgments. Both in the pretest and in the posttest, students provided item-specific performance judgments. After completing each test item, students had to indicate whether they thought their answer to the respective item was correct or not (yes/no). The single select option of the performance tests (the selected answer option could be right or wrong) represents a convergence between judgment and performance. By using one booklet for judgments and performance items, students had direct access to the items and their respective answers when judging the correctness of their answers. Students had five minutes to provide the 12/32 judgments in addition to the processing time of the pre- and posttest. Overall, 12 judgments were implemented in the pretest and 32 judgments in the posttest (Cronbach's $\alpha = 0.66$ for the pretest and $\alpha = 0.76$ for the posttest). According to Grier's (1971) nonparametric measure of response bias, students showed a low tendency to respond negatively on the judgments in pre- and posttest, $B''_{pre} = 0.38$ and $B''_{post} = 0.21$.

Preparation, post-processing, perceived usefulness of practice tests, and completion time. After the practice test phase was finished, students indicated how intensely they had prepared and post-processed the practice exams (i.e., "I intensely prepared for/post-processed the practice tests"). The answer scale was a 6-point Likert scale ranging from 1 "*not true at all*" to 6 "*absolutely true*". To understand how useful students perceived the practice tests regarding monitoring their performance, students who had participated in at least one practice test filled in a questionnaire scale consisting of 5 items (sample item: "The practice tests showed me my strengths and weaknesses"). Each item had to be answered on a 6-point Likert scale ranging from 1 "*not true at all*" to 6 "*absolutely true*" (Cronbach's $\alpha = 0.89$).

Based on recent research, we examined student completion time for the practice tests to investigate correlations between a process-based variable regarding the practice tests, performance on the posttest, and judgment accuracy scores (Tan et al., 2020). Students' completion time indicates whether students spent an appropriate amount of time on the practice tests to answer all items.

Table 1 Descriptive Statistics of all Variables (M, SD) for the Complete Samples of Studies 1 and 2

	Pretest *M (SD)*	Posttest *M (SD)*
Study 1		
Performance	62.81 (14.24)	63.22 (14.13)
Accuracy	63.85 (15.76)	58.69 (13.57)
Sensitivity	64.06 (25.84)	68.25 (19.10)
Specificity	55.62 (31.74)	34.26 (24.46)
Study 2		
Performance	32.66 (16.40)	79.98 (11.74)
Accuracy	62.31 (19.00)	77.48 (11.33)
Sensitivity	53.63 (32.68)	76.47 (17.55)
Specificity	64.07 (25.59)	52.26 (26.15)

All variables represent percentage scores

Data analyses

We recorded all scores into percentage scores. First, we calculated performance scores of the pre- and posttest. Second, to gain deeper insights into the relationship between participation in practice tests and judgment accuracy and to strengthen reliability and validity of the measurement of judgment accuracy (Rutherford, 2017), we calculated three scores: absolute accuracy as an absolute measure and sensitivity as well as specificity as relative measures of metacognitive accuracy.

The relations between repeated participation in practice tests and the criterion variables (posttest performance and posttest metacognitive scores) were studied via hierarchical regression analyses (H1, H3, Q1). In the first step, we inserted the respective pretest variables to control for potential differences in general performance level or judgment accuracy. That is, when investigating effects of repeated participation in practice tests on absolute judgment accuracy, absolute judgment accuracy regarding the pre-test was the predictor in the first step. In the second step, we studied the relation between the participation in repeated participation in practice tests as continuous variable and the criteria variables.

Results

Descriptive statistics of the main study variables are provided in Table 1. The number of taken practice tests did not significantly correlate with GPA ($r = –0.01$, $p = 0.88$) and was weakly related to the pretest score ($r = 0.15$, $p = 0.04$). That is, participation in retrieval practice does not seem to be strongly predetermined by a specific performance level. On average, students completed the practice tests in 12 min 9 s. The completion time did not correlate significantly with any other variable and was not part of the further analysis.

Table 2 displays preparation, post-processing, and perceived usefulness of practice tests for the subsample of students taking part in the practice tests. Students in Study 1 reported that they had not intensely prepared for the practice tests but that they had post-processed the practice tests intensely. Students participating in the practice test reported this was useful for monitoring their performance.

Table 2 Preparation, post-processing, and Perceived Usefulness of Practice Tests (M, SD) in Study 1 and Study 2

	Study 1 (pure practice tests) *M* (*SD*)	Study 2 (practice tests plus feedback) *M* (*SD*)
Preparation of practice tests	2.03 (1.13)	1.96 (1.13)
Post-processing of practice tests	4.20 (1.43)	4.92 (1.09)
Perceived usefulness of practice tests for monitoring	4.28 (1.11)	4.84 (0.74)
Perceived usefulness of feedback for monitoring	-	4.81 (0.83)

All variables were assessed on a 6-point Likert scale

Table 3 Regression Coefficients of Performance and Metacognitive Accuracy Scores on Pretest Variables and Practice Test Participation in Study 1(Upper Part) and Study 2 (Lower Part)

Step and predictor variable	Performance		Accuracy		Sensitivity		Specificity	
	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β
Study 1								
Step 1: Pretest variable	.06***	.25***	.00	.00	.04**	.20**	.03*	.17*
Step 2: Practice test participation	.07***	.27***	.03*	.18*	.03*	.17*	.02*	–.16*
Study 2								
Step 1: Pretest variable	.07**	.26**	.01	.08	.05*	.22*	.05*	.21*
Step 2: Practice test participation	.08**	.28**	.03	.17*	.06*	.24*	.00	–.05

* $p < .05$. ** $p < .01$. *** $p < .001$

Next, we regressed four variables separately (performance as well as three metacognitive accuracy scores) on students' frequency of their participation in practice tests.

Results in Table 3 indicate that controlling for prior knowledge, the number of practice tests students participated in was significantly related to the exam score. The more often students tested their knowledge, the better their exam score was.

Regarding accuracy, it seems that students who continuously engaged in practice tests provided more accurate judgments in the posttest. Similarly, students using more testing occasions showed higher sensitivity. That is, those students who repeatedly engaged in practice tests were able to detect more correct answers in the posttest. Conversely, students using more practice testing opportunities showed lower specificity. That is, with more testing students were less able to detect incorrect answers in the posttest.

Discussion

Study 1 indicates that students performed better in a real exam when taking practice tests (H1). Students who repeatedly self-tested their knowledge regarding the lecture topics might have elaborated more deeply on the learning content (Endres & Renkl, 2015; Fernandez & Jamet, 2016; Händel et al., 2020). In contrast to earlier studies, the current approach did not establish a control or placebo group with given alternate study strategies (Batsell et al., 2016; Fernandez & Jamet, 2016) but assessed the number of taken practice

tests as a continuous variable. Students could voluntarily participate in the practice tests but were not obliged to do so. Thereby, the current study depicts the beneficial effects of repeated retrieval in an authentic self-regulated learning situation.

In addition to the testing effect on performance, repeated participation in practice tests seems to transfer to judgment accuracy—but with differential effects for different accuracy scores (H3). In detail, students who engaged more frequently in practice tests were more accurate in their judgments (higher absolute accuracy) and were better able to detect correct items (higher sensitivity). This is a remarkable finding because, at the most, repeated practice test participation indirectly trained judgment accuracy without being accompanied by any additional training (not even repeated judging) on metacognitive awareness, as, for example, realized in other studies (Foster et al., 2017; Händel et al., 2020; Roelle et al., 2017). In other words, students did not only perform better but also seem to have generated a sense for their learning especially regarding their ability to detect correct answers. However, the negative relationship of repeated practice test participation with specificity indicates that this did not hold for incorrect items. On the contrary, students who repeatedly tested their knowledge might think themselves safe in the case of incorrect items. An explanation for the negative relationship with specificity is that students might be self-biased in the hope of passing the exam (Händel & Bukowski, 2019; Saenz et al., 2017; Serra & DeMarree, 2016) or they might be unaware of being unskilled to solve these particular items (Händel & Dresel, 2018). However, the slight negative response bias in the posttest should be interpreted as an indication that the first assumption does not hold.

Nevertheless, the results show that students have difficulty making accurate judgments and identifying incorrect responses. To exploit the full metacognitive potential of practice tests and to support students in assessing their performance, we implemented feedback on practice tests in the second study.

Study 2

Study 2 aimed to examine of the relationships between participation in practice tests, exam performance, and judgment accuracy when combined with individual corrective feedback (correctness of the answer and the correct solution).

Method

Study design

The design of Study 2 is similar to that of Study 1. The only difference was that students received item-specific individual feedback on the correctness of each answer and the correct answer immediately after finishing the practice tests. Students could save the items plus the corrective feedback after completing the practice test as a PDF file.

Sample

The participants were students of a psychology course for undergraduate students. There was no overlap with the course in Study 1 but the course took place in the same study term. Of the $N = 192$ students who took the corresponding exam, $n = 111$ participated in the pre- and posttest and provided item-specific judgments. The sample is large enough

to detect medium effects of $f^2 = 0.15$ in a regression analysis design with two predictors (presuming $\alpha = 0.05$, $1—\beta = 0.95$). Most of the participants were in the second year of their studies (88.3%). The majority of students were female (81.1%), which is typical for university introductory courses in this field of study. Students' GPA was similar to Study 1, $M = 2.54$, $SD = 0.54$.

Participation in the individual practice tests was as follows: 29 students participated in none of the practice tests, 29 students filled in one practice test, 15 students participated in two practice tests, 22 students participated in three practice tests, and 16 students took part in all four practice tests. Again, students were most likely to participate in the last practice test before the exam and participation in practice tests was treated as a continuous variable ranging from 0 to 4.

Instruments

The following sections give only a quick overview of instruments with a focus on aspects differing from Study 1 (see Study 1 for more information).

Prior knowledge (pretest). The prior knowledge covered 12 items with content from a previous mandatory psychology course, which most students took before the semester break.

Practice test. Each of the four voluntary practice tests was based on the course content of the respective previous three weeks: 1) clinical disorders, 2) motivation, 3) developmental psychology, and 4) social psychology. Immediately after finishing each practice test, students received individual item-specific feedback and had the opportunity to review their given answer as well as the sample answer of each item.

Test performance. The final course exam covered all lecture topics. Students had 50 min to solve the 36 (new to students) multiple-choice items (Cronbach's $\alpha = 0.75$).

Judgments. Both in the pretest and in the posttest, students were asked to provide item-specific performance judgments (item correct or not; Cronbach's $\alpha = 0.73$ for the pretest and $\alpha = 0.87$ for the posttest). Again, students showed a low tendency to respond negatively on the judgments in pre- and posttest, $B''_{\text{pre}} = 0.36$ and $B''_{\text{post}} = 0.36$.

Perceived usefulness of practice tests and feedback

Again, students indicated how intensely they had prepared and post-processed the practice exams and we assessed their completion time of the practice tests. In addition, students were asked to rate the usefulness of the practice tests and the usefulness of receiving feedback for monitoring their performance. Each scale consisted of the same five items, but referring to either the processing of the practice tests or the feedback provided after practice test completion (sample item:"The practice tests/The feedback on the practice tests showed me my strengths and weaknesses"). Each item had to be answered on a 6-point Likert scale ranging from "*not at all*" to "*completely*" (Cronbach's $\alpha = 0.88$ for the usefulness of the practice test and $\alpha = 0.93$ for the usefulness of the feedback).

Data analyses

As in Study 1, we examined the relationships of the number of taken practice tests, performance and judgment accuracy scores via hierarchical regression analyses (H2, H3, Q1). In addition, for the sample of students who participated in practice tests, we conducted paired sample *t*-tests to compare the perceived usefulness for monitoring one's performance due to taking practice tests versus the perceived usefulness for monitoring one's performance due to receiving feedback on the practice tests (Q3). Finally, to answer Q2 and Q3, we merged the two samples from Study 1 and Study 2. We used independent *t*-tests to compare the two samples from Studies 1 and 2 with respect to preparation and post-processing of practice tests as well as the usefulness of the practice tests for monitoring one's performance (Q2, Q3). This procedure, however, is only possible with regard to the self-reported variables of use and usefulness of practice tests because all other variables were assessed using instruments differing with regard to the respective course content.

Results

Table 1 provides descriptive statistics of the main variables in Study 2. The number of taken practice tests neither significantly correlated with GPA ($r = –0.17$, $p = 0.08$) nor with prior knowledge ($r = 0.17$, $p = 0.08$). Hence, participation in practice tests seemed to be independent of prior performance. On average, students completed the practice tests in 8 min 36 s. Again, we found no significant correlation with any other variable.

Table 3 illustrates the results of the hierarchical regression analyses. As in Study 1, we found a testing effect on exam performance. The more testing situations students used, the better their exam score was. In addition, a higher frequency of used practice tests resulted in better accuracy. Regarding the efficiency scores, the number of tests taken significantly predicted only sensitivity but not specificity. In detail, students who engaged in more practice tests showed higher sensitivity.

To check whether students perceived the test taking itself or the feedback as more helpful regarding monitoring their performance, we conducted a paired sample *t*-test with the two monitoring questionnaire scales (see Table 2 for descriptive values). Interestingly, students did not differ in their mean values regarding the perceived usefulness of practice test taking or the perceived usefulness of receiving feedback for monitoring their performance, $t(67) = 0.31$, $p = 0.76$.[1]

An independent sample *t*-test between Study 1 (pure practice tests) and Study 2 (practice tests plus feedback) showed that students who received individual corrective feedback (Study 2) rated the usefulness of participating in the practice tests as higher for monitoring their performance than students without individual corrective feedback in Study 1, $t(175) = –3.47$, $p < 0.001$, $d = 0.55$. The differences between the two groups of students can be considered as a medium effect. Finally, we compared students' preparation and post-processing of practice tests plus feedback with students' answers in Study 1 without feedback. As in Study 1, students did not intensely prepare for the practice tests; there was no significant difference between the two samples, $t(172) = 0.42$, $p = 0.67$, $d = 0.04$. However,

[1] Sample sizes are slightly lower because not all students who took part in the study filled in the questionnaire. In addition, only students who reported that they had saved the PDF file with the practice test items were considered in the analyses.

in Study 2, where students received individual corrective feedback, they post-processed the practice tests more intensely, which can be considered a medium effect, $t(175) = 3.66$, $p < 0.001$, $d = 0.57$.

Discussion

Study 2 mainly replicated the effects of Study 1. That is, the more frequently students participated in the practice tests, the better their performance on the posttest (H2), the higher their absolute accuracy, and the higher their sensitivity (H3) regardless of the completion time of the practice tests. However, the significant negative effect on specificity from Study 1 did not occur in Study 2 when providing feedback (H3). Probably, the corrective feedback informed students about their deficits and made them cautious (Raaijmakers et al., 2019). Still, although the negative relationship of engagement in practice tests and specificity did not occur for practice tests plus feedback, it needs to be noted that students who engaged in more practice tests plus feedback were not able to detect incorrect items better than students who engaged in less practice tests.

Interestingly, regarding the usefulness for monitoring performance of taking the practice tests on the one hand and receiving feedback on the other hand (Q2); we found no significant differences between the two questionnaire scales. However, this could also imply that students had difficulties in differentiating between pure practice tests and practice tests plus feedback as students immediately received the feedback (i.e., without any delay).

General discussion

The current two studies investigated participation in practice tests as a learning strategy via voluntary participation in practice tests. The studies focused on performance in a final exam as well as absolute and relative accuracy scores of metacognitive judgments. The analysis of different accuracy scores is a strength of this study and allows taking a close look at the relations between the amount of practice tests taken and measures of judgment accuracy. A valuable feature of both studies is that they examined undergraduate students' voluntary test taking behavior in an existing psychology course and that the design of the study can be considered ecologically valid.

Both with and without feedback, students who took practice tests not only performed better in a final course exam (H1, H2) but also showed higher absolute accuracy and higher sensitivity regarding their judgments (H3). Therefore, participation in practice tests as a learning strategy was related to students' ability to estimate personal performance, which is in line with previous classroom studies (Schwieren et al., 2017). Remarkably, students could improve their judgment accuracy by repeated practice test taking only, that is, without having practiced judgments. Beta weights of the regression analyses were of similar size across the two studies. This is an important finding as it shows that students showed higher accuracy after solely taking tests on course content. Practice tests participation itself might stimulate elaboration of the respective content and thereby establish a more solid knowledge base, which accordingly facilitates judgment accuracy (Barenberg & Dutke, 2018). This assumption is in line with the unskilled and unaware effect (Kruger & Dunning, 1999) revealing that students with higher topic knowledge provide accurate judgments. Moreover, similar results across the two studies regarding absolute accuracy and

sensitivity support findings by a recent study that found no difference in metacognitive accuracy whether feedback was provided or not (Raaijmakers et al., 2019).

By implementing two relative efficiency scores of judgment accuracy, we could show that the frequency of students' participation in practice tests significantly and positively predicts their sensitivity. Students seem to profit from the metacognitive cues of practice tests when they test their knowledge regularly during the semester with and without feedback. In case of lacking external feedback, it is likely that students self-generate internal feedback based on the cues during task performance (Koriat et al., 2008). An alternative explanation for the findings of both studies regarding sensitivity could be a student response bias. Students might more frequently indicate that the answer is incorrect in the pretest at the beginning of the semester and, based on better preparation, more frequently make positive judgments in the posttest at the end of the semester. In both studies, however, students tended to show a low negative response bias at both measurement times, so the assumption cannot be confirmed.

We found an interesting difference between the two studies regarding specificity (Q1). Focusing on pure practice tests, students who engaged in the tests were less able to detect incorrect answers in the posttest compared to students who did not frequently engage in practice testing. This was not the case in Study 2 with feedback. A possible assumption is that practice tests with feedback help students to monitor their *lacking* knowledge. That is, feedback on students' weaknesses might be eye opening with regard to one's knowledge and help accept one's deficits. This assumption is in line with findings by previous research showing positive effects of testing on retention of initially incorrect items (Vojdanoska et al., 2010; McDaniel et al., 2007). Nevertheless, the results of both studies suggest that educators who want to support students regarding their metacognitive judgment accuracy should consider strengthening student's specificity.

Additional questions on the preparation, post-processing and perceived usefulness of the practice tests provided further interesting insights into the comparability of the two studies (i.e., students did not intensely prepare for the ungraded practice exams) and on the specifics of the studies (Q2, Q3). Students who received individual corrective item feedback reported that they engaged more intensely in post-processing of the practice tests and reported a higher usefulness of the practice tests to monitor their performance than students engaging in pure practice tests reported. Hence, these results can be regarded as a first indicator that students experience and perceive their participation in practice tests with or without feedback differently—also with regard to its usefulness for performance monitoring. Feedback could be a tool to make the effective learning strategy more attractive to students who—so far—scarcely use this strategy (Blasiman et al., 2017; Tullis & Maddox, 2020).

Limitations and implications for future studies

While the two studies corroborate findings of previous studies and extend existing research designs with the analyses of several scores of metacognitive judgment accuracy, they do not allow drawing conclusions about the mechanisms behind the effects.

First, measurement issues might limit the study results. The studies did not control for guessing probability regarding the judgments. Actually, we assume that consideration of guessing probability would not compromise results in the chosen approach (dependent and independent variables would undergo the same correction in the regression analyses). Still,

the results on judgment accuracy need to be interpreted carefully and future studies might correct for guessing, for example, by implementing an additional response option regarding judgments (Vuorre & Metcalfe, 2021). Furthermore, future studies should implement a control condition. For example, as proposed by Greving and Richter (2018), with an additional control group, students could receive summaries for parts of the lecture (re-reading-condition), take practice tests for another part (testing condition), and receive feedback on chosen test items.

Second, to gain a deeper understanding of the underlying mechanisms, future studies should additionally assess process data and investigate differential feedback effects. Academic achievement and metacognitive monitoring are complex constructs that are naturally confounded with other variables of learning. Therefore, future studies should consider a process-based assessment of potential covariates. In specific, it is necessary to understand how students engage in learning besides the provided practice test and gain detailed insights information about students' preparation for the practice tests as well as for the final exam. Important indicators of students' learning engagement during the semester are the use of learning strategies as well as their learning time (Blasiman et al., 2017). For this purpose, we recommend the additional implementation of regular learning diaries together with the practice tests. Ideally, students should regularly participate in practice tests, use meaningful and effective learning strategies, and distribute their learning time evenly over the semester (Naujoks & Händel, 2020). Additionally, it is necessary to extend the knowledge about students' post-processing of the practice tests to understand how they regulate their learning after working on the tests. This might help to integrate the effects regarding performance and metacognitive judgment accuracy and the influence of feedback into the existing research literature, and thereby extend existing findings.

Another potential moderator for the effects of testing on performance and metacognitive judgment accuracy might be student motivation regarding participation in practice tests or additional learning activities. It seems reasonable, for example, that students with higher motivation participate in more practice tests, which in turn relates to better performance in the final exam. Current research by Tan et al. (2020) provides first insights that no such relationship between the number of practice tests and student motivation exists. In their study, only the completion time correlated positively with a performance-avoidance orientation of the students. The longer students took to complete the tests, the more likely were they to report avoidance aims regarding performance in the associated course. In the present study, students' completion time of the practice tests seemed adequate. Students reported intensive post-processing of the practice test and found them useful. We take this as a first indication that students engaged sufficiently with the practice tests and valued the implementation of such tests in regular university courses. Nevertheless, future studies should further investigate students' value of studying and their use of practice tests.

Third, in Study 2, the feedback informed students about the correct task solution, which goes beyond pure true/false feedback and thereby should foster students' engagement with the feedback (Fernandez & Jamet, 2016; Winstone et al., 2017). However, we collected no data on how, that is, with which strategies or frequency students further processed the feedback (Jönsson & Panadero, 2018). Admittedly, students might have only checked whether their practice test score would be sufficient to receive a specific grade. On the contrary, they might also have used the feedback information to correct misunderstandings or to become aware of metacognitive inaccuracies. Addressing other levels of feedback like direct feedback on judgment accuracy (Hattie & Timperley, 2007; Raaijmakers et al., 2019) or using prompts to encourage students' engagement with the feedback (Winstone et al., 2017) could strengthen the results.

Overall, our studies reveal that practice tests plus automatically generated individual corrective item feedback can be implemented easily via online tests. Hence, the results of the two current studies on performance and metacognitive judgment accuracy should encourage lecturers to provide students with practice tests. Our results on participation rates also indicate that educators might need to encourage students to steadily practice-test their knowledge.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics statement Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements.

Informed consent The participants provided their written informed consent to participate in this study.

Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy.

References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements, 11*(1), 159–177. https://doi.org/10.1037/h0093018

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876. https://doi.org/10.1002/acp.1391

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory and Cognition, 39*(1), 171–184. https://doi.org/10.3758/s13421-010-0002-y

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology. Applied, 24*(1), 43–56. https://doi.org/10.1037/xap0000133

Bangert-Drowns, R. L., Kulik, C.-L.C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Barenberg, J., & Dutke, S. (2018). Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory, 27*(3), 269–279. https://doi.org/10.1080/09658211.2018.1506481

Barenberg, J., & Dutke, S. (2021). Retrieval practice effects in a psychology lecture: Illustrating the relevance of study design, item difficulty, and selection of dependent measures. *Psychology Learning & Teaching*. https://doi.org/10.1177/14757257211049312

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2016). Ecological validity of the testing effect. *Teaching of Psychology, 44*(1), 18–23. https://doi.org/10.1177/0098628316677492

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory, 25*(6), 784–792. https://doi.org/10.1080/09658211.2016.1221974

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 229. https://doi.org/10.3389/fpsyg.2012.00229

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616. https://doi.org/10.3758/mc.36.3.604

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Chen, X., Zhang, M., & Liu, X. L. (2019). Retrieval practice facilitates judgments of learning through multiple mechanisms: Simultaneous and independent contribution of retrieval confidence and retrieval fluency. *Frontiers in Psychology*, *10*, 987. https://doi.org/10.3389/fpsyg.2019.00987

Cogliano, M. C., Kardash, C. A. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology, 56*, 117–129. https://doi.org/10.1016/j.cedpsych.2018.12.001

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: Knowing the test format enhanced metacognitive monitoring. *Metacognition and Learning, 5*(2), 195–206. https://doi.org/10.1007/s11409-010-9057-1

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Enders, N., Gaschler, R., & Kubik, V. (2021). Online quizzes with closed questions in formal assessment: How elaborate feedback can promote learning. *Psychology Learning & Teaching, 20*(1), 91–106. https://doi.org/10.1177/1475725720971205

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*, 1054. https://doi.org/10.3389/fpsyg.2015.01054

Fernandez, J., & Jamet, E. (2016). Extending the testing effect to self-regulated learning. *Metacognition and Learning, 12*(2), 131–156. https://doi.org/10.1007/s11409-016-9163-9

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 238–244. https://doi.org/10.1037/0278-7393.33.1.238

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9,* 2412. https://doi.org/10.3389/fpsyg.2018.02412

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*(6), 424–429. https://doi.org/10.1037/h0031246

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

Händel, M., & Bukowski, A.-K. (2019). The gap between desired and expected performance as predictor for judgment confidence. *Journal of Applied Research in Memory and Cognition, 8*(3), 347–354. https://doi.org/10.1016/j.jarmac.2019.05.005

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265–285. https://doi.org/10.1007/s11409-018-9185-6

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65,* 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jönsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. L. J. K. Smith (Ed.), *The Cambridge handbook of instructional feedback* (pp. 1–31). University Press.

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology, 59*(5), 251–257. https://doi.org/10.1027/1618-3169/a000150

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92*(4), 800–810. https://doi.org/10.1037//0022-0663.92.4.800

Koriat, A. (2019). Confidence judgments: The monitoring of object-level and same-level performance. *Metacognition and Learning, 14*(3), 463–478. https://doi.org/10.1007/s11409-019-09195-7

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In I. J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 117–135). Psychology Press.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 106–114. https://doi.org/10.1037/a0033699

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kubik, V., Jonsson, F. U., Knopf, M., & Mack, W. (2018). The direct testing effect is pervasive in action memory: Analyses of recall accuracy and recall speed. *Frontiers in Psychology*, *9*, 1632. https://doi.org/10.3389/fpsyg.2018.01632

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*(2), 94–97. https://doi.org/10.1177/0098628311401587

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. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology. Applied, 20*(1), 3–21. https://doi.org/10.1037/xap0000004

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education, 4*, 5. https://doi.org/10.3389/feduc.2019.0000

Naujoks, N., & Händel, M. (2020). Nur vertiefen oder auch wiederholen? Differenzielle Verläufe kognitiver Lernstrategien im Semester [Cram for the exam? Distinct trajectories of cognitive learning

strategy use during the term]. *Unterrichtswissenschaft, 48*(2), 221–241. https://doi.org/10.1007/s42010-019-00062-7

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning, 14*(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

Rivers, M. L., Dunlosky, J., & Joynes, R. (2019). The contribution of classroom exams to formative evaluation of concept-level knowledge. *Contemporary Educational Psychology*, *59*, 101806. https://doi.org/10.1016/j.cedpsych.2019.101806.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Chapter One - Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 1–36). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology, 109*(1), 99–117. https://doi.org/10.1037/edu0000132

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bullentin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction, 47*, 33–42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Saenz, G. D., Geraci, L., Miller, T. M., & Tirso, R. (2017). Metacognition in the classroom: The association between students' exampredictions and their desired grades. *Consciousness and Cognition, 51*, 125–139. https://doi.org/10.1016/j.concog.2017.03.002

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unawarein the classroom: College students' desired grades predict theirbiased grade predictions. *Memory & Cognition, 44*(7), 1127–1137. https://doi.org/10.3758/s13421-016-0624-9

Susser, J. A., & McCabe, J. (2012). From the lab to the dorm room: Metacognitive awareness and use of spaced study. *Instructional Science, 41*(2), 345–363. https://doi.org/10.1007/s11251-012-9231-8

Tan, T. Y., Jain, M., Obaid, T., & Nesbit, J. C. (2020). What can completion time of quizzes tell us about students'motivations and learning strategies? *Journal of Computing in Higher Education, 32*(2), 389–405. https://doi.org/10.1007/s12528-019-09239-6

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*(3), 429–442. https://doi.org/10.3758/s13421-012-0274-5

Tullis, J. G., & Maddox, G. B. (2020). Self-reported use of retrieval practice varies across age and domain. *Metacognition and Learning, 15*(2), 129–154. https://doi.org/10.1007/s11409-020-09223-x

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology, 24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09257-1

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and taxonoy of reciepience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538