Enhanced monitoring accuracy and test performance: incremental effects of judgment training over and above repeated testing

Marion Händel, Bettina Harder, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Händel, Marion, Bettina Harder, and Markus Dresel. 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.

Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing

Marion Händel[a,*], Bettina Harder[a], Markus Dresel[b]

[a] Department of Psychology, Friedrich-Alexander University Erlangen-Nürnberg, Germany
[b] Department of Psychology, University of Augsburg, Germany

Metacognitive processes play a major role in successful learning. One of these processes entails monitoring one's progress and correctness of acquired knowledge. Accurate monitoring has a significant impact on regulation activities and is assumed to lead to better performance (Kostons, van Gog, & Paas, 2012; Metcalfe & Finn, 2008; Nelson & Narens, 1990). Miller and Geraci (2014) emphasize that "the benefits of accurate metacognitive monitoring cannot be understated" (p. 139). For example, a student who recognizes knowledge gaps or difficulties in understanding will ideally invest more resources, such as time-on-task or effort, and implement appropriate learning strategies in order to understand and retain the topic. Conversely, a student who is convinced of already understanding the content will not engage in further learning, which might lead to low performance if the student does not actually possess the required knowledge (Winne & Perry, 2000). Hence, accurate monitoring is considered to have important implications for learning success. However, even experienced learners such as university students appear to struggle with accurately judging their performance level and generally overestimate personal performance (Foster, Was, Dunlosky, & Isaacson, 2017; Händel & Fritzsche, 2016).

Although it seems crucial to assist students in accurate monitoring, there is only limited evidence on how to successfully accomplish this. Previous training attempts were either only effective in laboratory settings or improved only monitoring accuracy but not performance. Some procedures focused on repeated monitoring judgments but did not take into account the effect of repeatedly practicing content on monitoring (Bol, Hacker, O'Shea, & Allen, 2005). In addition, these studies neglected to provide students with feedback on task performance and judgment accuracy, which is necessary to correct misunderstandings and to assemble knowledge (Bol et al., 2005; Miller & Geraci, 2014). The current study aimed to overcome these limitations. For this purpose, metacognitive training with several features (psychoeducation, repeated testing and judging, and feedback) was developed. Its incremental effects on monitoring accuracy and performance over and above those achieved by repeated testing plus feedback were investigated in an authentic classroom setting. In addition, the study examined the interindividual development of overconfidence over the course of a study term.

1. Theoretical background

Metacognitive monitoring is addressed in both models of metacognition and models of self-regulated learning (Nelson & Narens, 1990; Winne, 1996). During monitoring, learning outcomes are compared with previously set benchmarks in order to evaluate whether set goals

* Corresponding author.
E-mail address: marion.haendel@fau.de (M. Händel).

have been met. Nelson and Narens (1990) described the interplay of monitoring and regulation: Monitoring processes are relevant because only they can depict inconsistencies or knowledge gaps that require further learning and regulation of learning. Monitoring is always related to a specific criterion with which information is compared and accordingly provides information about self-perceived knowledge for a concrete task (Winne, 1996).

1.1. Metacognitive judgments and their accuracy

Metacognitive judgments can be characterized as a tool used to explicitly undertake monitoring and to measure the outcome of monitoring processes. They can vary according to several characteristics (for an overview, see Nelson & Narens, 1990). For example, judgments can be made before or after testing (pre-versus postdictions) and on a global or local level. Global judgments are assessed at the level of the whole test and are driven by more superficial cues such as domain-familiarity or self-concept. In contrast, local or item-specific judgments refer to each single item and can be based on task-specific cues.

Most students seem to have difficulties in judging their performance accurately. A couple of studies in educational settings clearly indicated students' overconfidence regarding exam performance (Foster et al., 2017; Fritzsche, Händel, & Kröner, 2018; Miller & Geraci, 2011). Especially low-performing students provide overconfident judgments and are accordingly considered to be *unskilled and unaware* (Händel & Fritzsche, 2016; Kruger & Dunning, 1999). Two main reasons are discussed as being responsible for students' inaccurate judgments. On the one hand, students may not be able to provide judgments that are more accurate. That is, students might not have recourse to item-specific cues (including item difficulty, ease of processing, or ability to explain meaning) that would help them to evaluate the adequacy of their judgments (Koriat, Nussinson, Bless, & Shaked, 2008; Kruger & Dunning, 1999; Thiede, Griffin, Wiley, & Anderson, 2010). On the other hand, students may not be motivated to make more accurate judgments (cf. Roelle, Schmidt, Buchau, & Berthold, 2017). For example, students might not really want to put much effort into making accurate judgments because they do not understand the value of doing so (Gillström & Ronnberg, 1995). In addition, students' judgments might be driven by motivational influences such as wishful thinking (Händel & Bukowski, 2019; Serra & DeMarree, 2016).

1.2. Training approaches

In educational settings, the two assumptions of why students are inaccurate judges of own performance convert into different training approaches. First, students can be made aware of the *relevance of monitoring accuracy*, for example, by providing them with knowledge about it. Second, in a context with abundant practice and feedback opportunities, a further promising approach is exercising metacognitive judgments (de Bruin & van Gog, 2012). *Practice and feedback* should point to diagnostic cues that help to judge personal performance accurately. Here, it needs to be distinguished whether the focus is (a) on practicing content with the aim to influence judgments implicitly or (b) on explicitly practicing judgments.

1.2.1. Knowledge about the relevance of judgments

Students may only be motivated to put effort into accurate judgments when they are aware of their importance and the negative consequences of inaccurate judgments. Hence, informing students about the dangers of making overconfident judgments seems a promising intervention. This was studied by Roelle et al. (2017) in three computer-based experimental studies indicating that informing students about the danger of overconfidence increased their monitoring activities, led to more cautious judgments, and fostered the acquisition of conceptual knowledge (medium effects: $\eta^2 \leq 0.11$). Since the studies were experimental with quite a short learning phase, it needs to be clarified whether the results transfer to authentic classrooms.

A classroom training study with fifth-graders by Huff and Nietfeld (2009) combined approaches of strategy instruction and practicing judgments. Students were required to think about the importance of confidence judgments and to evaluate their judgments regarding under- or overconfidence. This significantly influenced monitoring scores ($\eta^2 \leq 0.14$) but not reading performance. In line with Roelle et al. (2017), the lack of training effects might be explained by the young age group's low knowledge regarding regulation strategies. Overall, educational input seems a promising approach, which, however, might need to be combined with other features to be beneficial in a classroom setting.

1.2.2. Retrieval practice

Testing can easily be embedded in classrooms—not only for summative assessment but also as a learning tool—and seems beneficial in several regards. Previous research has found that repeated retrieval positively influences future learning and recall, which is known as the *testing effect* or *repeated retrieval effect* (Adesope, Trevisan, & Sundararajan, 2017; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006). In their meta-analysis, Adesope et al. (2017) found that the testing effect was robust across different types of item formats, samples, settings, or retention intervals. While most of the studies used experimental designs, the testing effect also occurred in authentic learning settings with meaningful learning material (Endres & Renkl, 2015; McDaniel et al., 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011). Usually, research in this tradition used identical stimuli for learning and repeated testing—yet an important question for educational settings is whether taking one exam benefits students with regard to a future exam containing different questions. Some studies using within-subject designs replicated the testing effect for non-identical items in the repeated testing situations or the final test (A. C. Butler, Black-Maier, Raley, & Marsh, 2017; McDaniel, Wildman, & Anderson, 2012). For example, quizzes with feedback enhanced exam performance in two classroom-based experiments, even when the quiz items differed from the exam items ($d \leq 0.83$; McDaniel et al., 2012).

It has been suggested that repeated testing also influences metacognitive monitoring (Tullis, Finley, & Benjamin, 2013). For example, Roediger and Karpicke (2006) found that repeated testing in contrast to repeated studying of learning materials had a medium effect on students' predictions of future recall in a final exam ($\eta^2 = 0.06$). In a recent study, undergraduate students in a repeated testing group were less overconfident and had higher test performance compared to students without testing ($d \leq 1.27$; (Fernandez & Jamet, 2016). Studies using within-subject designs also found higher performance, higher accuracy and less overconfidence for practice-tested items compared to non-tested items in a text-learning setting ($d \leq 0.67$; Barenberg & Dutke, 2018) as well as in a regular undergraduate course setting ($\eta^2 \leq 0.85$; Cogliano, Kardash, & Bernacki, 2019). Hence, there is growing evidence that repeated retrieval does not only influence performance but also metacognitive accuracy. Barenberg and Dutke (2018) proposed two explanations for the potential testing effect on metacognitive judgments. First, retrieval practice might serve as an implicit cue (Kelemen, 2000). That is, students can use information from earlier retrieval attempts like ease of processing as a knowledge-based cue to judge the outcome of a task response. Second, retrieval might stimulate elaboration of the respective content and thereby lead to a more solid knowledge base. Students' elaborate knowledge and the availability of information should accordingly enhance monitoring accuracy.

1.2.3. Practicing judgments

So far, some studies have investigated the influence of practicing judgments on metacognitive accuracy and performance. Those that were conducted in settings with high external validity, however, either lack (appropriate) control groups (Bol et al., 2005; Callender, Franco-Watkins, & Roberts, 2015; Foster et al., 2017), which limits

interpretation of results, or failed to achieve training effects (Bol et al., 2005; Foster et al., 2017; Hacker, Bol, & Bahbahani, 2008; Kelemen, Winningham, & Weaver, 2007).

Several reasons can be assumed as to why training procedures failed. First, when predictions rather than postdictions are implemented, students cannot base their judgments on elaborate knowledge before test-taking and, hence, have little chance to improve (see Bol et al., 2005). This might have been the case in a study, in which students predicted their exam performance across 13 exams yet did not become more accurate judges over time but remained overconfident (Foster et al., 2017). Second, the implementation of judgments and feedback at the overall test level might be responsible for the lack of both training effects and developmental changes in judgment accuracy over time (Foster et al., 2017; Hacker, Bol, Horgan, & Rakow, 2000). Hence, when providing global judgments, students have limited access to relevant cues. In addition, global judgments might not inform students about the specific tasks or content they should study further (van Loon & Roebers, 2017). Item-specific judgments seem to be more informative for students, but only rarely have these been investigated. A laboratory study, however, indicates that repeatedly made item-specific judgments (without being tested on the items) led to a comparable effect on final performance ($d = 0.57$) as repeated testing (Jönsson, Hedner, & Olsson, 2012).

Finally, some studies cannot clarify whether training effects result from repeated testing or from repeated judging (Barenberg & Dutke, 2018; Bol et al., 2005). To separate the effects, Kelemen et al. (2007) used a design with three groups: a group that practiced vocabulary plus judgments (item-specific predictions for a future cued-recall test), one that practiced vocabulary only, and a no-practice group. In the posttest session, students of all groups practiced vocabulary, provided predictions, and completed a cued-recall test. Practicing in general lowered overestimation compared with no-practice, but the additional practice of judgments did not lower overestimation compared with practicing vocabulary only. However, again only predictions were used and students received no feedback, which calls for more intense interventions with feedback on the item-level (D. L. Butler & Winne, 1995; Labuhn, Zimmerman, & Hasselhorn, 2010). For example, medium to large effects for monitoring ($\eta^2 = 0.10$) and performance ($\eta^2 = 0.16$) were found in a study that combined monitoring exercises with (classroom) feedback (Nietfeld, Cao, & Osborne, 2006).

In sum, the empirical evidence obtained so far suggests that monitoring accuracy can potentially be improved. However, previous studies have several limitations. First, most of the studies implemented global judgments only. Such judgments neither rely on relevant cues nor are informative for students' future learning processes. Second, the studies often used only one accuracy measure, which differed across studies. Third, they often lacked a control group or did not take the initial judgment level into consideration, which limits the interpretation of group differences. Fourth, the potential of training regarding performance, especially in studies with high external validity, was either not investigated or not fully exploited. Finally, there is little evidence on the development of monitoring accuracy over time.

1.3. Aims of the study

We aimed to develop a training procedure that is beneficial for students' monitoring accuracy and performance in an authentic classroom setting. We based our training procedure on the two deficiencies that were assumed to be responsible for inaccurate judgments (lacking skills and lacking will). We implemented item-specific judgments because they were assumed to be informative for students, allow task-specific feedback on monitoring accuracy to be provided, and are considered superior to global judgments from a methodological perspective as well (Dunlosky & Lipko, 2007).

Specifically, our study pursued three main goals. First, with respect to the testing effect, we studied whether repeated testing in an authentic setting with content-parallel non-tested items helps students to not only perform better but also to judge performance more accurately (measured via absolute and relative scores of monitoring accuracy) compared with students without repeated testing. Second, we tested for an additional effect of metacognitive training compared with repeated testing only. Third, we aimed to depict interindividual developments of judgments.

Our hypotheses regarding the testing effect (H1) and the additional effect of metacognitive training (H2) were as follows:

H1a. Students engaged in repeated testing show more accurate metacognitive judgments in a final exam than students without repeated testing.

H1b. Students engaged in repeated testing show higher performance in a final exam than students without repeated testing.

H2a. Students engaged in repeated testing plus metacognitive training show more accurate metacognitive judgments in a final exam than students engaged in repeated testing only.

H2b. Students engaged in repeated testing plus metacognitive training show higher performance in a final exam than students engaged in repeated testing only.

Concerning the development of metacognitive judgments, our hypothesis was as follows:

H3. On average, students are overconfident in their judgments and their overconfidence decreases over the course of the metacognitive training.

2. Method

2.1. Procedure

The quasi-experimental study was implemented in the ecologically valid setting of regular psychology courses for undergraduate teacher training students and lasted over the period of one study term (13 weeks). All students were enrolled in a lecture; in the first week, a pretest was conducted and in the last week, the final exam (posttest) took place. Both pretest and posttest consisted of knowledge tests, accompanied by item-specific metacognitive judgments. During the term, students had the opportunity to participate in one of two content-identical courses accompanying the lecture (Courses A or B, same lecturer, each 90 min per week, and referring to lecture topics). Students chose course participation before the term started with respect to their weekly schedules. Students were not informed previously that the courses would differ in any way and, thus, could not have known about the different treatments within the courses. Students from Course A were assigned to the *metacognitive training group and* students from Course B to the *testing group*. Students who did not participate in any of the courses but only in the pre- and posttest formed the *control group*. Table 1 presents an overview of the implemented procedures in the three conditions. In the testing and metacognitive training group, mock exams were implemented as repeated testing occasions. They represent low-stakes exams not affecting students' final grades.

2.1.1. Testing group

Throughout the term, students in the testing group participated in three mock exams, covering content of the preceding three sessions. One week after each mock exam, students received individual written feedback on their task performance: For each item, students were provided with the correct answer, their given answer, and the respective correctness of their answer.

2.1.2. Metacognitive training group

The metacognitive training combined three training elements, namely psychoeducation, practicing judgments, and feedback on judgement accuracy. In the first course session, students received information

Table 1
Overview of the study procedure per group.

Week	Measures and interventions	Group		
		Control	Testing	Metacognitive training
1 (pretest)	Prior knowledge test + judgments	x	x	x
	Psychoeducation			x
4	Mock exams		x	x
	Judgments			x
5	Feedback on task performance		x	x
	Feedback on judgments			x
8	Mock exams		x	x
	Judgments			x
9	Feedback on task performance		x	x
	Feedback on judgments			x
11	Mock exams		x	x
	Judgments			x
12	Feedback on task performance		x	x
	Feedback on judgments			x
13 (posttest)	Final exam + judgments	x	x	x

Note. In addition to the intervention, students regularly participated in the weekly courses lasting 90 min each.

on the significance of monitoring accuracy, adapted from Roelle et al. (2017). Because students are usually overconfident, this psychoeducational approach informed students about overconfident judgments and asked them to elaborate on their danger and potential consequences. Students' comprehension of the psychoeducational input was confirmed by a test with open- and closed-ended items on its content (possible scores of 0–4, $M = 3.88$, $SD = 0.40$). Throughout the term, students in the metacognitive training group worked on the three identical mock exams as students in the testing group. However, the test items were accompanied by item-specific judgments. One week after each mock exam, students received a sheet with their original answers and feedback on their judgment accuracy in addition to feedback on their task performance (i.e., whether their answer to an item was correct or not and whether they detected it as such in their judgment). Accordingly, students were provided with one of the following four feedback options per judgment: (a) "correct: your answer was true", (b) "correct: your answer was wrong", (c) "incorrect: you were overconfident", and (d) "incorrect: you were underconfident". To see how students dealt with the individual feedback, they were explicitly asked about this at the end of the term. The majority of students in the metacognitive training group paid attention to the feedback: 90% of the students indicated that they analyzed the feedback on task performance, and 62% indicated that they studied the feedback regarding their judgments.

2.2. Sample

Based on the medium effect sizes reported in previous studies, a power analysis was conducted. To detect medium effects of $\eta^2 = 0.10$ in an ANCOVA design (presuming $\alpha = .05$, $1 - \beta = 0.95$), a minimum sample size of 143 participants (48 per group) was needed. After providing study consent, 221 students voluntarily participated in study. All students were informed about the possibility to participate in mock exams. Six students were excluded because they changed courses during the term. Six further students were excluded because they provided metacognitive judgments for less than two thirds of the items in the posttest. The final sample thus consisted of $N = 209$ students with a gender distribution typical for German teacher education programs, with 78.9% female (Schmidt, 2013). Students' final high school grade (ranging from 1-very good to 6-insufficient) was 2.42 ($SD = 0.49$) and can be regarded as average. Across all three groups, the majority of students were in their first ($n = 150$) or second ($n = 48$) year of studies (the remaining students in their third or fourth year). Sample sizes were as follows: $n = 47$ students in the control group, $n = 54$ students in the testing group, and $n = 108$ students in the metacognitive training group—the large metacognitive training group may be a result of the subjectively more attractive weekday of the respective course. The majority of the students (80.6%) in the testing and metacognitive training group attended at least two of the three mock exams (no significant group differences in course attendance rates; $p = .46$). As an incentive, six gift cards (equivalent to € 10) were raffled off, three for participation in the pretest and three for participation the posttest. In addition, for regular participation in the testing and metacognitive training group, students received a USB flash drive (equivalent to € 5).

2.3. Measures

2.3.1. Performance tests

The pretest and the three mock exams consisted of 12 items each. The pretest covered content from a previous mandatory psychology course. The mock exams each covered the topics of the preceding three lecture sessions. The posttest (final course exam with 36 items) tested the content of the complete lecture—using different items than in the pretest and mock exams. Hence, items in all tests were presented only once. All tests were multiple-choice and curricularly valid. For each item, students had to choose one out of four possible answers. A sample item was "The statement 'I am gifted for the subject I study' expresses the person's … (a) academic self-concept, (b) mastery-approach goal, (c), learning goal, (d) attribution to internal, variable causes." The implemented tests spanned several topics, including motivation, intelligence, developmental psychology, clinical disorders, and social psychology. Accordingly, the tests cannot be considered homogenous one-dimensional tests. This is especially true for the mock exams, which were not developed with regard to a measure of performance but served as an opportunity for students to assess understanding of course content; McDonald's ω consequently indicated fairly low values with 0.62, 0.51, and 0.56. Internal consistency for the pre- and posttest was satisfying with $\omega = 0.68$ and 0.73. The test scores of the pre- and the posttest were significantly related with each other ($r = 0.33$, $p < .001$).

2.3.2. Metacognitive judgments

The mock exams in the metacognitive training group and the pretest and posttest for the whole sample were each accompanied by item-specific metacognitive judgments. After completing each test item, students had to indicate whether they thought their answer to the respective item was correct or not (yes or no). Metacognitive judgment items were printed in the same booklet as the performance items so that students had direct access to the items and their respective answers when judging the correctness of their answer. McDonald's ω of the metacognitive judgments across the five tests was .92, .81, 0.90, 0.92, and 0.90.

For all tests, absolute and relative accuracy scores were calculated (Bol & Hacker, 2012). Bias reflects the degree of underconfidence (negative bias values) or overconfidence (positive bias values) and is computed as the signed differences between performance p_i and judgments c_i, averaged over the n items (Schraw, 2009):

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(c_i - p_i) \quad (1)$$

Bias ranges from −1 (underconfidence) to 1 (overconfidence). Absolute accuracy as the absolute value of bias indicates the fit of performance and judgment. Scores close to zero point to accurate monitoring and values close to one indicate inaccurate judgments:

$$\text{Absolute accuracy} = \frac{1}{n}\sum_{i=1}^{n}|c_i - p_i| \quad (2)$$

These two absolute scores provide unique information. For instance,

when a student over- and underestimates item performance in equal shares, only absolute accuracy provides substantial information and points to the student's low metacognitive accuracy. Conversely, if a student solves all items correctly but assumes all are incorrect, bias provides additional information aside from absolute accuracy and indicates students' underconfidence.

In line with previous studies, the sensitivity and specificity scores were calculated (Rutherford, 2017; Schraw, Kuch, & Gutierrez, 2013). Sensitivity indicates the relative frequency of accurately detected correct answers:

$$\text{Sensitivity} = \frac{\sum \textit{items accurately detected as correct}}{\sum \textit{correct items}} \quad (3)$$

Specificity indicates the relative frequency of accurately detected incorrect answers:

$$\text{Specificity} = \frac{\sum \textit{items accurately detected as incorrect}}{\sum \textit{incorrect items}} \quad (4)$$

These relative scores take into account the number of items solved correctly and incorrectly and are therefore superior to the signal detection categories of hits, false alarms, misses, and correct rejections, which depend on total test performance (Green & Swets, 1966). The two scores complement each other: Sensitivity provides diagnostic information about the detection of correct items; specificity refers to incorrect items. Hence, a student can have a high sensitivity and a high specificity at the same time.

2.4. Data analyses

To investigate training effects (Hypotheses H1 and H2), group differences in posttest variables (monitoring scores and performance) as dependent variables were analyzed. Due to significant group differences in performance and monitoring scores in the pretest (Pillai's trace $p < .001$),[1] analyses of covariance with group as a between-subject factor and respective pretest variables as a covariate were conducted. That is, for group differences in posttest bias, for example, pretest bias was used as a covariate.[2] Planned contrasts tested for differences between the testing group and the control group and between the metacognitive training group and the testing group.

Moreover, for students in the metacognitive training group who continuously practiced judgments, the development of judgment bias was investigated (H3). We focused on judgment bias because psychoeducation explicitly aimed to reduce overconfidence. Latent growth curve modeling (LGCM) was used to fit the individual development of judgment bias; the full information maximum likelihood (FIML) approach was used to deal with missing data. For students who participated in the respective mock exams, the nonresponse missing rate ranged from 1.3% to 5.5%.

3. Results

For descriptive statistics of the pretest and posttest measures, see Table 2. Jittered points boxplots per group are displayed in Fig. 1 to illustrate the distributions of all dependent variables corrected for pretest values.

ANCOVAs are reported to reveal group differences in the posttest variables. Significant and medium to large training effects were found for bias ($F(2, 205) = 9.10$, $p < .001$, $\eta^2 = 0.08$), absolute accuracy ($F(2, 205) = 17.15$, $p < .001$, $\eta^2 = 0.14$), and specificity ($F(2, 204) = 3.45$, $p = .033$, $\eta^2 = 0.03$). In addition, a significant and large effect for exam performance was found ($F(2, 205) = 23.63$, $p < .001$, $\eta^2 = 0.19$). No significant group difference emerged for sensitivity ($F(2, 203) = 0.34$, $p = .72$, $\eta^2 = 0.00$).

To examine these main effects further, planned contrasts tested the effect of pure testing (H1) and the additional effect of metacognitive training (H2) on monitoring and performance (Table 3). In comparison to the control group (no testing), repeated testing with feedback on task performance in the testing group resulted in significant and medium to large effects for absolute accuracy and performance. As expected, students in the testing group developed lower absolute accuracy scores, that is, better monitoring accuracy, than students in the control group (H1a). Moreover, repeated testing led to a higher final exam score than that of the control group (H1b).

Above and beyond the effects of pure testing plus individual feedback on task performance, the metacognitive training positively influenced several monitoring scores and performance. Planned contrasts indicated significant differences for bias, absolute accuracy, specificity, and performance between the testing group and the metacognitive training group (H2a). Specifically, students in the metacognitive training group less strongly overestimated their performance in the final exam (lower bias), provided more accurate judgments (lower absolute accuracy scores), and were better at detecting incorrect answers in the final exam (higher specificity). Finally, the planned contrast regarding final exam performance indicated a significant and medium effect: The metacognitive training led to higher performance in the final exam than the testing group (H2b).

Lastly, the development of judgment bias in the metacognitive training group was studied, Fig. 2 illustrates this development at the group level. Overall, students were initially overconfident. Subsequently, bias decreased, with students even overregulating (bias score below zero in mock exams two and three), but ending up at nearly zero in the final exam. To account for possible individual differences in the development of judgment bias, LGCM was applied. A latent basis model (Grimm, Ram, & Hamagami, 2011) indicated good fit (CFI = 1.00, RMSEA = 0.054, $\chi^2(7) = 7.06$, $p = .42$; see Fig. 3 for the model specification). The estimated mean intercept differed significantly from zero. Concurrently, significant intercept variance points out substantial interindividual differences in their—overall overconfident—starting levels. The slope factor revealed a significantly negative mean but non-significant interindividual variance. Hence, students lowered their overconfidence across mock exams but showed negligible interindividual differences in their developmental trajectories, that is, a uniform development in bias (H3). Accordingly, the non-significant intercept-slope-correlation indicates that the development of judgment bias was independent of individual starting levels.

4. Discussion

The current study implemented training that was highly effective not only for monitoring accuracy—assessed via several scores—but also for final exam performance. The study combined relevant training features in an authentic learning setting. To support students' skills and will to monitor their understanding of study content accurately, we informed them about the importance of monitoring accuracy, repeatedly provided them with mock exams over the course of the term, and asked them to judge their performance for every single exam item. Moreover, students received detailed individual feedback regarding their performance and judgments.

To our knowledge, the present study is the first to investigate the additional effects of metacognitive training in comparison to the effects of repeated testing, which was in turn compared with a control group without any practice opportunities. Further strengths of the study are that monitoring accuracy was investigated via several metacognitive measures

[1] Post-hoc tests revealed significant group differences in the pretest variables only in comparison with the control group but not between the testing and metacognitive training group.

[2] We furthermore conducted all analyses with pretest performance as an additional covariate to consider its potential influence. This did not substantially alter the results. Similarly, correcting for guessing, which was possible only for performance but not for monitoring scores, did not alter the results.

Table 2
Means (standard deviations) of the pretest and posttest variables per group.

Group	Bias		Absolute accuracy		Sensitivity		Specificity		Performance	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Control group	0.17 (0.31)	0.13 (0.26)	0.44 (0.18)	0.38 (0.12)	0.62 (0.30)	0.78 (0.22)	0.51 (0.33)	0.40 (0.31)	0.35 (0.18)	0.57 (0.11)
Testing group	0.18 (0.24)	0.10 (0.20)	0.34 (0.17)	0.30 (0.10)	0.76 (0.28)	0.83 (0.18)	0.46 (0.34)	0.39 (0.31)	0.44 (0.20)	0.66 (0.11)
Metacognitive training group	0.15 (0.24)	0.02 (0.17)	0.40 (0.18)	0.28 (0.09)	0.68 (0.33)	0.80 (0.16)	0.45 (0.32)	0.48 (0.29)	0.47 (0.21)	0.71 (0.10)
Total	0.16 (0.26)	0.05 (0.21)	0.39 (0.18)	0.31 (0.10)	0.69 (0.31)	0.80 (0.18)	0.47 (0.33)	0.44 (0.30)	0.44 (0.20)	0.67 (0.12)

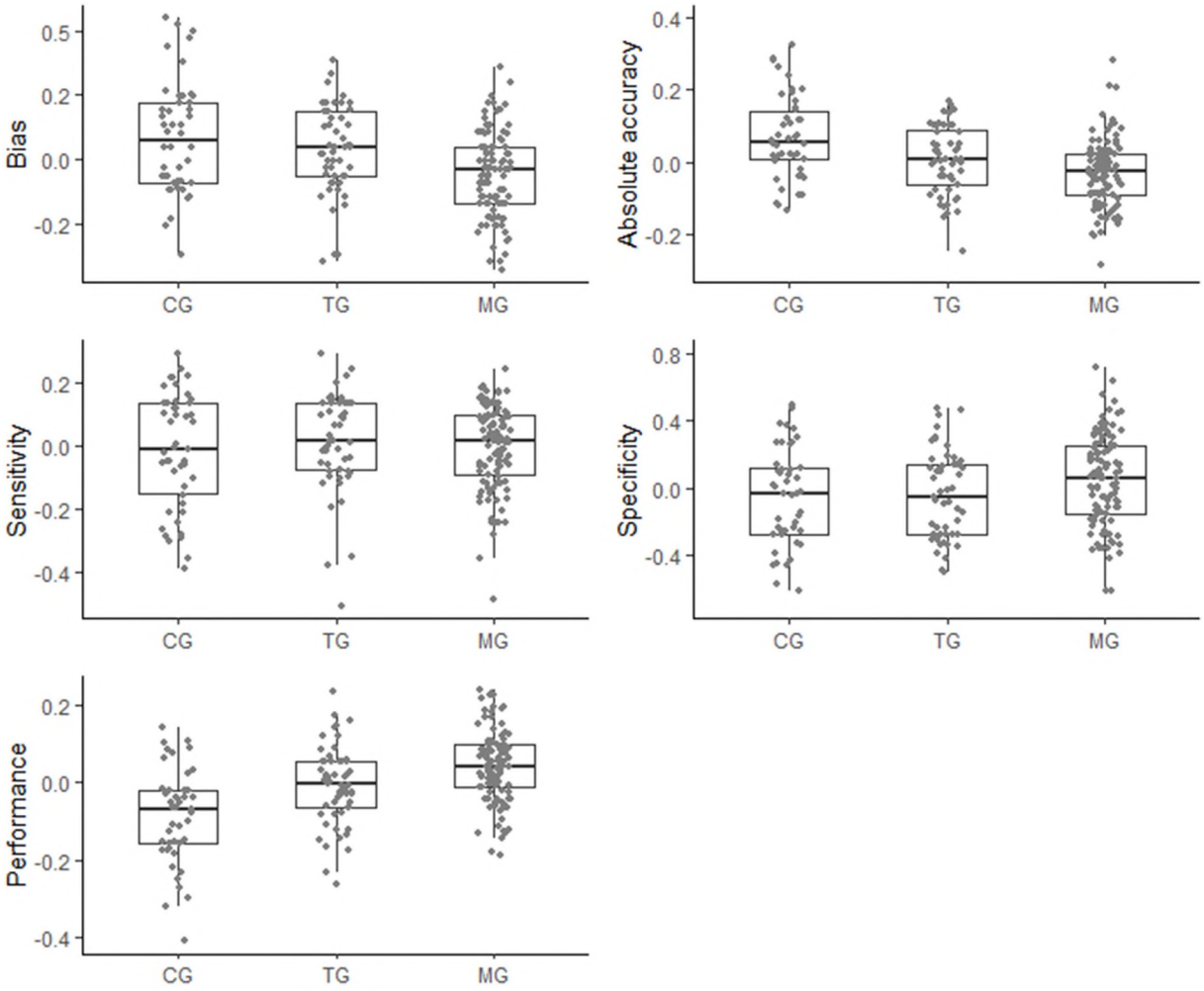

Fig. 1. Multi-group boxplots for the residuals of dependent posttest variables regressed on pretest variables, separately displayed for the control group (CG), testing group (TG), and metacognitive training group (MG).

Table 3
Group comparisons (planned contrasts) for the posttest scores.

Group comparison	Bias		Absolute accuracy		Sensitivity		Specificity		Performance	
	p	*g*	*p*	*g*	*p*	*g*	*p*	*g*	*p*	*g*
Testing group vs. control group	.287	0.15	.001	0.74	.218	0.29	.905	0.05	.001	0.77
Metacognitive training group vs. testing group	.003	0.53	.027	0.26	.216	−0.22	.018	0.32	.002	0.52

Note. Hedges' *g* is calculated as an effect size measure. Its interpretation is similar to Cohen's *d* (it uses pooled weighted standard deviations instead of pooled standard deviations).

for item-specific judgments in order to depict a more holistic pattern of training effects (Boekaerts & Rozendaal, 2010; Schraw et al., 2013). Finally, our study substantially contributes to the understanding of the development of monitoring judgments—LGCM were used as a state-of-the art procedure to consider individual variability in this development.

4.1. Repeated testing improves not only performance but also monitoring accuracy

Overall, significant and strong effects of repeated testing with unknown items were found. Repeated testing enhanced performance in an

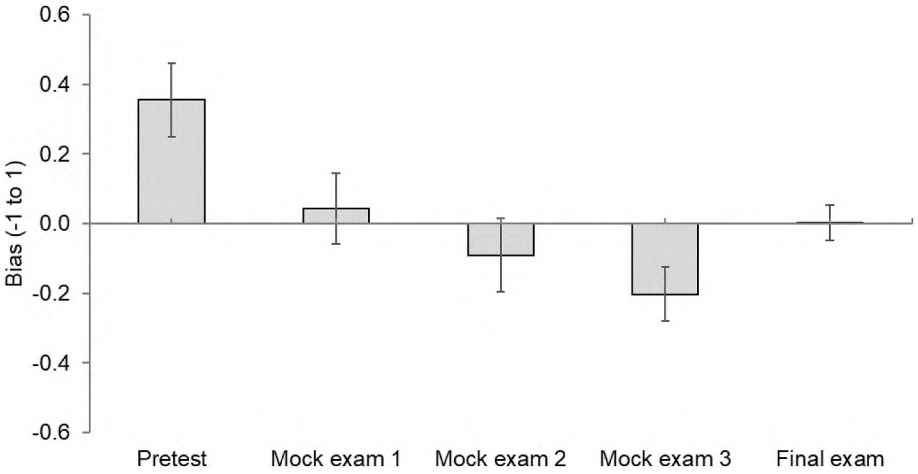

Fig. 2. Judgment bias across exams for students who received metacognitive training. Error bars represent 95% confidence intervals.

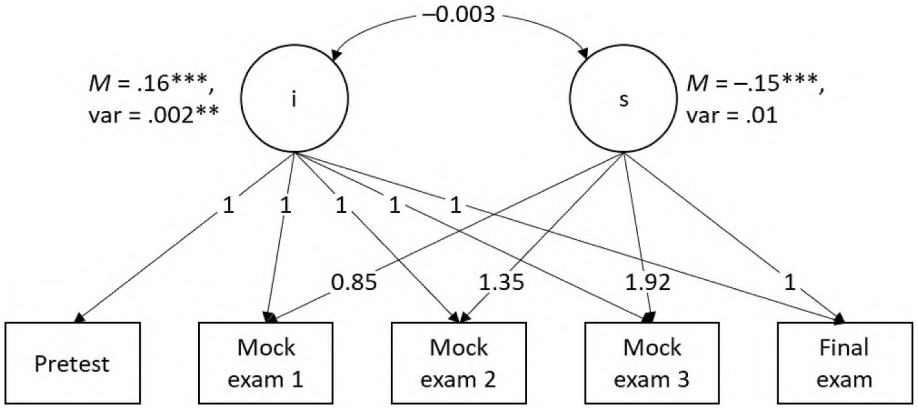

Fig. 3. Latent basis growth curve model with a latent intercept (i) and slope (s) factor for judgment bias across mock exams.

authentic classroom context (H1b), which is in line with previous findings in laboratory and classroom settings (McDaniel et al., 2007; Roediger & Karpicke, 2006). The implementation of mock exams in regular classrooms was beneficial for students' exam performance. Because students had prior experiences with multiple choice exams, dealing with the format of the test items alone could not have been responsible for the testing effect. Instead, the examination of content is suggested to lead to better understanding and retention. This is in line with a previous study, in which mental effort as an indicator for semantic elaboration contributed to the testing effect on performance (Endres & Renkl, 2015).

Furthermore, repeated testing was also beneficial for one out of four measures for monitoring accuracy, namely absolute accuracy (H1a). Students who engaged in repeated testing might not only have elaborated on the content but also on how well they understood the content. Altogether, students seem not only to achieve more elaborate knowledge but also seem to be better able to recognize whether they correctly solved the items in the final exam or not. Previous research has provided evidence for positive training effects on monitoring when the same items were repeatedly tested and judged (Barenberg & Dutke, 2018; Jönsson et al., 2012). Our research substantially contributes to the field by expanding the testing effect on metacognitive monitoring to material that differed for each testing situation—an important finding relevant to classroom settings (Kelemen et al., 2007). Retrieving content in authentic learning settings seems to serve as an implicit cue to reflect upon item correctness and thereby leads to higher monitoring accuracy.

Further research is needed to understand why repeated testing helped students to be more accurate but not, for example, to be less overconfident. Think-aloud protocols or open responses seem to be useful methods for revealing the cues to which students refer (Dinsmore & Parkinson, 2013; Hacker et al., 2008; Händel & Dresel, 2018; Thiede et al., 2010). This should help to understand the basis and purpose of students' judgments.

4.2. Metacognitive training adds to the effects of repeated testing

While repeated testing plus feedback already resulted in strong effects on performance and on absolute accuracy, the metacognitive training significantly added to the effects. The combination of psychoeducation, which provided students with knowledge on the relevance of accurate judgments, with several judgment practice opportunities plus feedback on their judgment accuracy positively influenced metacognitive monitoring (significant group differences for three out of four monitoring accuracy scores—H2a) and performance (H2b). This is a remarkable result, especially with regard to previous studies, which were not or only partly successful in supporting students to become more accurate judges and better achievers (Bol et al., 2005; Kelemen et al., 2007; Miller & Geraci, 2011). Specifically, students less strongly overestimated their final exam performance, provided more accurate judgments, and were better able to detect incorrect items. Only correct items were not better identified as such. This might be explained via the psychoeducational approach, which mainly focused on the dangers of overconfidence (Roelle et al., 2017). Accordingly, students might have focused more on incorrect items than on correct ones. Hence, sensitivity and specificity might have been affected by psychoeducation and feedback differently (cf. Schraw et al., 2013). In sum, the metacognitive training influenced monitoring accuracy as a proximal training variable and final exam performance as a distal variable. Especially the

combined procedure of psychoeducation together with practice and feedback opportunities might have supported students in engaging in monitoring and regulation activities.

4.3. Development of judgment bias: students adjust their judgments over time

Investigating judgment bias in the metacognitive training group in more detail revealed interesting findings. Group means at a descriptive level indicated that students were initially overconfident, which is in line with our expectations and earlier research results (Händel & Fritzsche, 2016; Kruger & Dunning, 1999). Subsequently, students lowered their bias, and even temporarily provided underconfident judgments, which corresponds to the underconfidence-with-practice effect (Koriat, Sheffer, & Ma'ayan, 2002). The metacognitive training seemed to lead students to overregulate their judgments. Remarkably, however, in the final exam, students no longer showed strong over- or underconfidence. Hence, the received feedback on being underconfident in the respective items of the previously taken mock exams might have compensated for the over-regulation of judgment bias in the final exam.

A LGCM fitted this nonlinear development of judgment bias with significant estimates for intercept and slope, indicating interindividual differences in the pretest and quite similar development across students (H3). The early decrease in overconfidence after the pretest (where no feedback was provided) points to an influence of the psychoeducation focusing on the negative consequences of being overconfident (Roelle et al., 2017). The later decrease in bias was most likely due to the feedback on judgments.

4.4. Limitations and directions for future research

The limitations of this study are that it cannot reveal how dedicated the students of all three groups were in their preparations for the final exam. This is especially true for the control group, which did not take part in any of the mock exams. Similarly, no information was collected about how students prepared for the mock exams (these were not graded) or how seriously they took the judgments in the final exam (in which the item solution but not the judgments were graded). Nevertheless, previous studies have not revealed any differences in judgment accuracy due to extra credit for accurate judgments (Hacker et al., 2008; Miller & Geraci, 2011). In addition, the study yields limited information about how extensively students elaborated on the feedback that was provided one week after answering the related items. Because students in our sample might not have realized the relevance of feedback on judgments, it can be assumed that metacognitive training effects might even be boosted, for example, when prompting students how to deal with the feedback.

A further limitation relates to sample sizes, which differed between the testing and the metacognitive training group. Students' schedules are thought to be the main reason for sample size differences. While access to the experimental groups was available to all students (via an online platform) and all provided information regarding the courses was identical (duration, lecturer, and content) when students enrolled, experimental groups differed in scheduled time, which by chance led to higher participation rates in the metacognitive training group.

Furthermore, results might be biased due to pretest differences between the testing and the metacognitive training group, on the one hand, and the control group on the other hand. Control group students might not only have had schedule difficulties but might not have deemed it necessary to attend an accompanying course. Although pretest scores were used as covariates, it cannot completely be ruled out this led to a certain overestimation of the effects of the testing group. It is important to note that this could not have impaired the additional effects of the metacognitive training above and beyond repeated testing—the central innovative finding of the present study—that were found.

Finally, we cannot completely pinpoint whether the metacognitive training effects and the development of judgment bias are based on the combination of psychoeducation and repeated judging plus feedback or on single training features. Students might have been prompted by the psychoeducation on the dangers of overconfidence and students might have benefitted from practicing judgments and thereby considering item-specific cues (Koriat et al., 2008; Thiede et al., 2010) and respective feedback in order to derive according judgments.

Of relevance for future research is to unravel which information and cues students take into consideration when providing judgments, especially after receiving individual feedback on performance and judgments (Callender et al., 2015; McDaniel et al., 2007; Miller & Geraci, 2011). Metacognitive judgments are presumed to influence performance not directly but rather via subsequent control processes (Nelson & Narens, 1990). Students might have used learning approaches that helped them to better understand and retain the course content relevant for the final exam. Possible regulation activities include that students put in more study effort or adapted their learning strategies. Subsequent studies should investigate not only monitoring but also regulation, and especially the influence of their interplay on performance. That is, an important further step is to study which strategies students implement when they notice that they lack knowledge and whether, for example, this depends on the type of knowledge or task (Endres & Renkl, 2015).

5. Conclusions

Due to the relevance of accurate monitoring during learning and because students are usually inaccurate and overconfident judges of their own performance, successful interventions to improve monitoring and performance are needed. This study indicates that the facilitation of practicing content during a study term supported students' performance and monitoring accuracy. In addition, the combination of making students aware of the relevance of accurate monitoring and providing them with judgment practice opportunities and respective feedback was found to be highly successful in terms of monitoring accuracy and performance. These findings should encourage future classroom interventions to support students in becoming accurate judges of their own performance. In addition, the study contributed to the understanding of interindividual developments of judgments over time. Interindividual differences in the intercept point to the relevance of considering (meta-)cognitive prerequisites when studying the development of judgment accuracy. Further research is needed to provide evidence on the interplay of monitoring judgments and performance and on the (mediating) role of regulation activities. A next step toward this understanding consists of exploring which monitoring activities trigger which regulation strategies, and how these regulation strategies influence students' performance.

Acknowledgements

This research was supported by a grant from the Emerging Talents Initiative (ETI) of the Friedrich-Alexander University of Erlangen-Nuremberg, grant number 2018/1_Phil_02. We thank Tobias Debatin for fruitful discussion of latent growth curve modelling.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

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*, 659–701. https://doi.org/10.3102/0034654316689306.

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

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*, 372–382. https://doi.org/10.1016/j.learninstruc.2009.03.002.

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

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*, 269–290. https://doi.org/10.3200/JEXE.73.4.269-290.

de Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*, 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003.

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology, 23*, 433–446. https://doi.org/10.1037/xap0000142.

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

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2015). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*, 215–235. https://doi.org/10.1007/s11409-015-9142-6.

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.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension. *Current Directions in Psychological Science, 16*, 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology, 6*, 1–6. 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*, 131–156. https://doi.org/10.1007/s11409-016-9163-9.

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*, 1–19. https://doi.org/10.1007/s11409-016-9158-6.

Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning, 13*, 159–177. https://doi.org/10.1007/s11409-018-9182-9.

Gillström, A., & Ronnberg, J. (1995). Comprehension calibration and recall prediction accuracy of texts: Reading skill, reading strategies, and effort. *Journal of Educational Psychology, 87*, 545–558. https://doi.org/10.1037/0022-0663.87.4.545.

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

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development, 82*, 1357–1371. https://doi.org/10.1111/j.1467-8624.2011.01630.x.

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*, 101–121. https://doi.org/10.1007/s11409-008-9021-5.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170. https://doi.org/10.1037//0022-0663.92.1.160.

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*. 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*, 265–285. https://doi.org/10.1007/s11409-018-9185-6.

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition, 44*, 229–241. https://doi.org/10.3758/s13421-015-0552-0.

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*, 161–176. https://doi.org/10.1007/s11409-009-9042-8.

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*, 251–257. https://doi.org/10.1027/1618-3169/a000150.

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

Kelemen, W. L., Winningham, R. G., & Weaver, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*, 689–717. https://doi.org/10.1080/09541440701326170.

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

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162. https://doi.org/10.1037//0096-3445.131.2.147.

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*, 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004.

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*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*, 173–194. https://doi.org/10.1007/s11409-010-9056-2.

van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology, 31*, 508–519. https://doi.org/10.1002/acp.3347.

McDaniel, M. A., Anderson, J.s. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 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*, 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174–179. https://doi.org/10.3758/PBR.15.1.174.

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

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition, 29*, 131–140. https://doi.org/10.1016/j.concog.2014.08.008.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–141. https://doi.org/10.1016/S0079-7421(08)60053-5.

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. https://doi.org/10.1007/s10409-006-9595-6.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology, 17*, 382–395. https://doi.org/10.1037/a0026252.

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

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*, 99–117. https://doi.org/10.1037/edu0000132.

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.

Schmidt, G. (2013). Hochschulen in Bayern [Universities in Bavaria]. *Bayern in Zahlen. Fachzeitschrift für Statistik, 7*, 395–402.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 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.

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

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*, 331–362. https://doi.org/10.1080/01638530902959927.

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*, 429–442. https://doi.org/10.3758/s13421-012-0274-5.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*, 327–353. https://doi.org/10.1016/S1041-6080(96)90022-9.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.). *Handbook of self-regulation* (pp. 531–566). San Diego: Academic Press.