Original Article

7- and 8-Year-Olds' Struggle With Monitoring

Inaccuracy Persists Despite Feedback

Kristin Kolloff, Claudia M. Roebers, and Florian J. Buehler

Institute of Psychology, Faculty of Human Sciences, University of Bern, Switzerland

Abstract: An often-replicated finding in metacognition research is that children overestimate their performance. To date, only a few studies have investigated the possible effects of item-specific feedback on metacognitive monitoring in young children. This study examined whether first-graders benefit from feedback to improve metacognitive monitoring discrimination when completing a paired-associates task for consistency. Over six sessions, *N* = 112 children evaluated whether they solved tasks correctly or not and gave item-specific confidence judgments. One group obtained only performance feedback; the other group received additional feedback on whether their performance matched their monitoring judgments. Results revealed that children could adequately discriminate between correct and incorrect answers in their confidence judgments. However, neither type of feedback improved metacognitive monitoring discrimination. We discuss the results against the theoretical background of Efklides' self-regulation model and the cue utilization approach.

Keywords: metacognition, monitoring, confidence judgments, feedback, course of intervention

Metagkognitive Überwachung bei 7 – 8-Jährigen. Ungenauigkeit bleibt trotz Feedback bestehen

Zusammenfassung: Ein oft replizierter Befund in der Metakognitionsforschung ist die Selbstüberschätzung von Kindern bezüglich der eigenen Leistung. Bisher haben nur wenige Studien die möglichen Auswirkungen von itemspezifischem Feedback auf die metakognitive Überwachung bei jungen Kindern untersucht. Wir untersuchten, ob Erstklässler bei der Bearbeitung einer Paarassoziationslernaufgabe von Feedback profitieren, um die metakognitive Überwachung zu verbessern. In sechs Sitzungen bewerteten *N* = 112 Kinder, ob sie Aufgaben richtig oder falsch lösten und gaben itemspezifische Sicherheitsurteile ab. Eine Gruppe erhielt nur Leistungsfeedback, die andere Gruppe erhielt Leistungsfeedback und Feedback über die Übereinstimmung zwischen Leistung und Sicherheitsurteil. Die Ergebnisse zeigten, dass Kinder in ihren Sicherheitsurteilen adäquat zwischen richtigen und falschen Antworten unterscheiden können. Jedoch konnten beiden Formen von Feedback die Genauigkeit der metakognitiven Überwachung nicht verbessern. Die Ergebnisse werden vor dem theoretischen Hintergrund von Efklides' Selbstregulationsmodells sowie dem Cue-Utilization-Ansatz diskutiert.

Schlüsselwörter: Metakognition, Überwachung, Sicherheitsurteil, Feedback, Interventionsverlauf

Children and adults experience uncertainty in many everyday life situations. For example, when trying to remember shopping lists, departure times of trains, or school assignments (like learning the European capitals), individuals may feel certain about some information, while they feel less certain about other information. The ability to differentially and accurately judge one's confidence, ranging from being totally uncertain to being very certain, has proven to be associated with, among others, academic outcomes for children, adolescents, and adults (Dunlosky & Metcalfe, 2009). Given its importance for school achievement in children, it is not surprising that practitioners and researchers alike seek means to improve the accuracy of these so-called *metacognitive monitoring judgments* (Dignath et al., 2008; Nelson & Narens, 1990). Researchers have broadly investigated how monitoring accuracy can be increased sustainably (e.g., Kelemen et al., 2007; van Loon & Roebers, 2020). Feedback, not only on the performance but also on one's metacognitive monitoring judgments, has the potential to promote monitoring accuracy, which, in turn, may facilitate self-regulated learning (Efklides, 2008). Feedback is thought to be critical for calibrating monitoring. However, the impact of feedback experiences on young children's metacognitive monitoring accuracy needs to be examined in more detail (Efklides & Metallidou, 2020). We can assess the accuracy of children's monitoring using a variety of measures: Measures of absolute monitoring accuracy provide information on whether the learner is overconfident, underconfident, or well-calibrated (Pajares & Miller, 1997; Schraw, 2009), whereas measures of relative monitoring accuracy quantify the ability to differentiate correct from

incorrect responses (Dunlosky & Metcalfe, 2009). The present approach includes absolute and relative monitoring accuracy. Previous research showed that, independent of the measure used, children's metacognitive monitoring judgments seem rather inaccurate (Dunlosky & Lipko, 2007; Foster et al., 2017). Children and adults typically exhibit high overconfidence in their abilities (absolute accuracy; Destan & Roebers, 2015; Dunlosky & Rawson, 2012), while relative monitoring accuracy appears to be an earlier emerging skill that nevertheless undergoes a process of further calibration. In the present contribution, we describe and analyze data from a metacognitive training study conducted with 7- to 8-year-old students. Over six weeks, we provided feedback on item-specific performance and/or monitoring accuracy to explore whether and in what respect children's monitoring accuracy may benefit from repeated feedback.

We integrate two theoretical frameworks here, which serve as the background for the present approach: Efklides' model of self-regulated learning (SRL; 2011) for the broader developmental and educational perspective; and Koriat's cue-utilization framework (1997) for addressing the microprocesses at work during each task. Regarding the SRL model, Efklides (2006) emphasized the role of metacognition and its building blocks, with metacognitive experiences (particularly monitoring), metacognitive knowledge, and metacognitive control interacting with aspects of the task (e.g., task difficulty) and motivation (e.g., a feeling of comprehension). Efklides (2006) describes metacognitive experiences as being aware of metacognitive processes while completing a task. For example, when metacognitive experiences give an accurate picture of task mastery, this leads to adequate metacognitive control decisions (such as restudy decisions), enabling efficient SRL and increasing motivation. However, if an individual judges a task to be easy (because it looks familiar) while it is not, metacognitive experiences and control decisions are likely to be faulty and motivation decreases (frustration because the task was not fully accomplished sets in). By gaining experiences over time and within a task, metacognitive experiences are thought to undergo a process of calibration. That is, the involved monitoring processes are assumed to become increasingly accurate in attuning to actual performance. Thus, feedback is thought to be critical for the calibration process (Efklides & Metallidou, 2020).

Regarding this calibration process, Koriat's (1997) cue-utilization framework assumes that, based on experiences (e.g., during learning and recall), individuals come to discover so-called heuristics (i.e., mnemonic rules of thumb) and use them as cues for their monitoring. According to Koriat (1997), individuals consider different cues when monitoring. Thus, intrinsic cues are characteristics of learning tasks that are related to recall from memory (e.g., judgment of how easily something is learned). Extrinsic cues arise from the learning environment in which learning occurs (e.g., the number of times an item was studied). Mnemonic cues are internal indicators that infer how well an item has been learned. Differences in retrieval fluency, ease of recall, or familiarity serve as subjective cues for metacognitive monitoring, with familiar and supposedly easily remembered information typically receiving higher confidence ratings and vice versa (Koriat et al., 2009). In other words, if an individual tries to remember the names of European capital cities in a test situation, some might come easily to mind, while others require longer thinking. Some heuristics negatively bias monitoring (large font size is often thought to lead to better memory than small font; Mueller et al., 2014). But, generally, when individuals have made numerous subjective experiences with internal indicators (valid mnemonic cues), their monitoring benefits and cues become more important (Hertzog et al., 2002). That is, cues become fine-tuned and adjusted with increasing experience and through feedback, leading to improved monitoring accuracy over time and with increasing age (Koriat, 1997; Koriat & Levy-Sadot, 2001; Roebers et al., 2019; van Loon & Roebers, 2020).

Entering school increases children's opportunities to discover and use cues and engage in the above-mentioned fine-tuning of their monitoring. Previous research has shown that even 3- to 4-year-olds can metacognitively distinguish between incorrect and correct recall (emerging relative monitoring accuracy). From the age of 5 to 6 years onwards, they rudimentarily use heuristics (e.g., Geurten et al., 2017; Gonzales et al., 2022). Yet, young children's metacognitive monitoring is often faulty and best characterized by overconfidence regarding the correctness of single answers and overall performance (e.g., Dapp & Roebers, 2021; Destan & Roebers, 2015; Lipko et al., 2009). Although relative and absolute monitoring accuracy increases with age and experience, metacognitive control decisions have been found to rely heavily on overoptimistic monitoring and insufficient metacognitive discrimination, even in 4th graders (Bayard et al., 2021).

Against the theoretical background, a combination of instruction in school, task assignments, and personal and formal feedback should initiate and foster the development of accurate monitoring (Dunlosky & Metcalfe, 2009). Efklides (2008) posits that monitoring accuracy is affected not only by cognitive processes at the individual level but also by explicit information at the social level. Corrective feedback is one form of social interaction that may guide children's awareness of metacognitive processes. A review by Hattie and Timperley (2007) found that feedback is most effective when it is oriented toward

the learning process rather than focusing exclusively on the performance outcomes. Without structured intervention, inaccurate monitoring becomes the rule rather than the exception, even in university students (Dunlosky & Rawson, 2012). At the same time, however, giving concrete feedback proved to be effective in improving monitoring accuracy, at least in adults (Miller & Geraci, 2011; Nietfeld et al., 2006). The most common type of feedback in educational settings is *performance feedback* (Butler & Winne, 1995), which informs learners whether a task was correctly solved. A few studies assessed the effect of global performance feedback on children's overestimation, but, for the most part, this has not been effective (Geurten et al., 2017; Lipko et al., 2009; Shin et al., 2007). The value of feedback is enhanced by also taking inaccurate monitoring into account (Hattie & Timperley, 2007; van Loon & Roebers, 2021).

We argue that, unlike performance feedback, *metacognitive feedback* comprises multiple cues derived from repeated metacognitive experiences of mastering the task. Furthermore, from the social perspective, personal feedback about the current state of one's metacognitive monitoring processes may enhance the association between mnemonic cues and thoughts about judgments while processing a task (Efklides, 2006, 2011; Koriat, 1997). Only two studies addressed the influence of item-specific feedback on monitoring accuracy in young children, revealing that giving item-specific feedback on errors (performance feedback), giving item-specific monitoring feedback (van Loon & Roebers, 2020), or observing an adult model (Lipko-Speed et al., 2018), can significantly affect young children's recognition of errors or their recognition of ignorance (*relative monitoring accuracy*). Yet, in these studies, the children remained overconfident despite repeated feedback and recognized only a maximum of one-third of their errors (i. e., *absolute monitoring accuracy*). Against the background of findings proving young children's use of heuristics, these findings nevertheless suggest that – at least in principle – young children can adjust their monitoring accuracy after receiving item-specific feedback, with feedback serving most likely as a task- and item-specific metacognitive experience. Presently, it remains unclear whether feedback can affect children's metacognitive ability to discriminate between correct and incorrect responses. It also remains unclear what kind of feedback – and specifically whether performance or monitoring feedback – is most effective. Metacognitive discrimination, however, is an essential prerequisite for the detection and correction of errors as well as for efficient SRL.

The present study uses measures of relative and absolute monitoring accuracy to examine the impact of feedback on the development of children's monitoring accuracy. Empirical evidence has demonstrated emerging metacognitive skills that gradually develop from the age of 4 onwards. As a replication of previous research, we therefore expected that children at the age of 7 would be able to discriminate significantly between incorrect and correct responses (relative monitoring accuracy). One of the main objectives concerned whether relative monitoring accuracy increases throughout metacognitive training by providing two qualitatively different types of feedback (performance feedback vs. monitoring feedback). Various monitoring experiences, based on item-specific feedback, provide mnemonic cues that not only enhance cue validity but also influence future metacognitive monitoring judgments (Hertzog et al., 2002; Koriat, 1997). We thus hypothesized that the ability to discriminate between correct and incorrect answers would significantly increase in both conditions over the course of training. Regarding the second objective, we addressed the effectiveness of one of the two feedback qualities. According to Efklides (2011), repeated metacognitive experiences enrich children's knowledge of their own metacognitive skills. The interaction between metacognitive experiences with different aspects of task accomplishment could lead to improved calibration of monitoring accuracy in subsequent monitoring judgments. In light of this, we hypothesized that in the *Monitoring Feedback* condition the discrimination ability would increase more than in the *Performance Feedback* condition. Concerning the third objective, previous studies have shown that, despite attempts to improve absolute monitoring accuracy, children remain persistent in their overconfidence (Kelemen et al., 2007; Lipko et al., 2009). Therefore, we hypothesized that, despite repeated feedback, the children in both feedback conditions would maintain their overconfidence (absolute monitoring accuracy).

Method

The data presented here are part of a larger project designed as a pretest/training/posttest study. We used the data from a six-week training period for the analyses in this contribution, employing a different paired-associates memory-learning task in the pretest and posttests. During the training, we used a 4-point confidence scale, whereas in the pretest and posttest, we used a 7-point confidence scale. These differences ensured that the posttest was a near-transfer compared to the training tasks. Pretest and posttest data are published separately since the main study investigated a different research question (the pre-registration for the main study can be found: https://osf.io/mwnsy). Data collection took place

between April and June 2021. None of the tasks were part of the official school curriculum. The study received approval from the faculty's Ethics Committee and was conducted in accordance with the Declaration of Helsinki (Faculty of Humanities of the University of Bern; approval number: 2020–10–00005).

Participants

The final sample consisted of $N = 112$ children between 7 and 8 years of age (55% girls, M_{age} = 7 years, 4 months, SD = 4.8 months, range = 6 years, 8 months to 8 years, 6 months) from seven public schools in the German speaking part of Switzerland. We randomly assigned $n = 51$ children (45.5%) to the Performance Feedback and $n = 61$ (54.5%) children to the Monitoring Feedback condition. All students were regularly enrolled in the first grade and sufficiently fluent in German. Their parents gave written informed consent. Before each session, the children were asked for oral assent and informed that they could terminate at any point during the study. One child rejected participation and was excluded from the study. For the analyses, it was necessary to have data for both correct and incorrect recognition; for this reason, we excluded participants from analyses separately for every session if they had items either all incorrectly recognized ($n_{t1} = 8$, $n_{t2} = 0$, $n_{t3} = 8$, $n_{t4} = 1$, $n_{t5} = 3$, $n_{t6} = 8$) or all correctly recognized ($n_{t1} = 0$, $n_{t2} = 3$, $n_{t3} = 1$, $n_{t4} = 7$, $n_{t5} = 1$, $n_{t6} = 2$). This resulted in slightly varying degrees of freedom in the analyses reported below.

Procedure

We integrated the paired-associates task for consistency into a cover story where two protagonists helped to feed animals at a zoo (each session with a different class of species and their habitats: insects, rodents, birds, etc.). The children completed the sessions individually, with each session lasting for about 20 min, together with 5–9 peers in a quiet room at their school. We presented the tasks on a tablet computer with a touch screen (10.4") and headphones. Two trained experimenters responsible for the technical preparation also closely supervised children in the rare case that they needed help with the tablet computer or headphones.

Materials and Measures

We used a paired-associates learning task with multiple-choice recognition. It has been shown that variations of paired-associates learning tasks provide successful results (e.g., Buehler et al., 2021; Destan et al., 2017). In the first session, the participants received detailed explanations about the task and were instructed to remember the picture pairs (animals and their preferred food) and provide confidence judgments about their choice in the recognition test (four answer alternatives). A 4-point smiley scale was used to collect their confidence judgments following the recognition test (see Figure 1). Confidence judgment scales between two and seven anchors have been successfully used with primary school children (Dapp & Roebers, 2021; Roderer & Roebers, 2010; Tsalas et al., 2015). We selected items with varying difficulty from a pool of items to ensure sufficient variability. The item difficulty index (i.e., the percentage of students who answered an item correctly) ranged from .90 (easy items) to .11 (difficult items); the items were presented in a randomized order. Posthoc item analyses revealed the tasks consisted of 16% easy, 71% medium, and 13% difficult items. Internal consistency for the paired-associates learning task confidence judgments within each session was acceptably reliable (8 items per session: $\alpha_{t1} = .72$, $\alpha_{t2} = .79$, $\alpha_{t3} = .75$, $\alpha_{t4} = .77$, $\alpha_{t5} = .76$, $\alpha_{t6} = .79$).

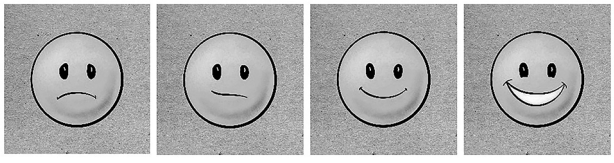

Figure 1. 4-point confidence judgment scale.

Memory Task

Before the test phase started, each participant completed a practice trial to familiarize themselves with the task. The task consisted of five different phases: learning phase, delay phase (filler task), recognition phase, memory-monitoring phase (confidence judgments), and feedback phase, depending on the experimental condition (see Figure 2; performance feedback or monitoring feedback).

In the learning phase, we presented 12 sets of pair-associated pictures consecutively for 5 s in each session, showing every item pair on a single slide that consisted of an animal and its preferred food.

In the delay phase, the children completed a simple 1-minute filler task to prevent memory strategies.

In the recognition phase, one animal appeared on the left-hand side and four pictures of food on the right-hand side of the screen. One picture of food was correct, whereas the other three alternatives appeared as food for other animals and were, thus, familiar to the participants. Before the participants started the recognition test, we implemented four observational learning trials. Based on four different item pairs, the two protagonists demons-

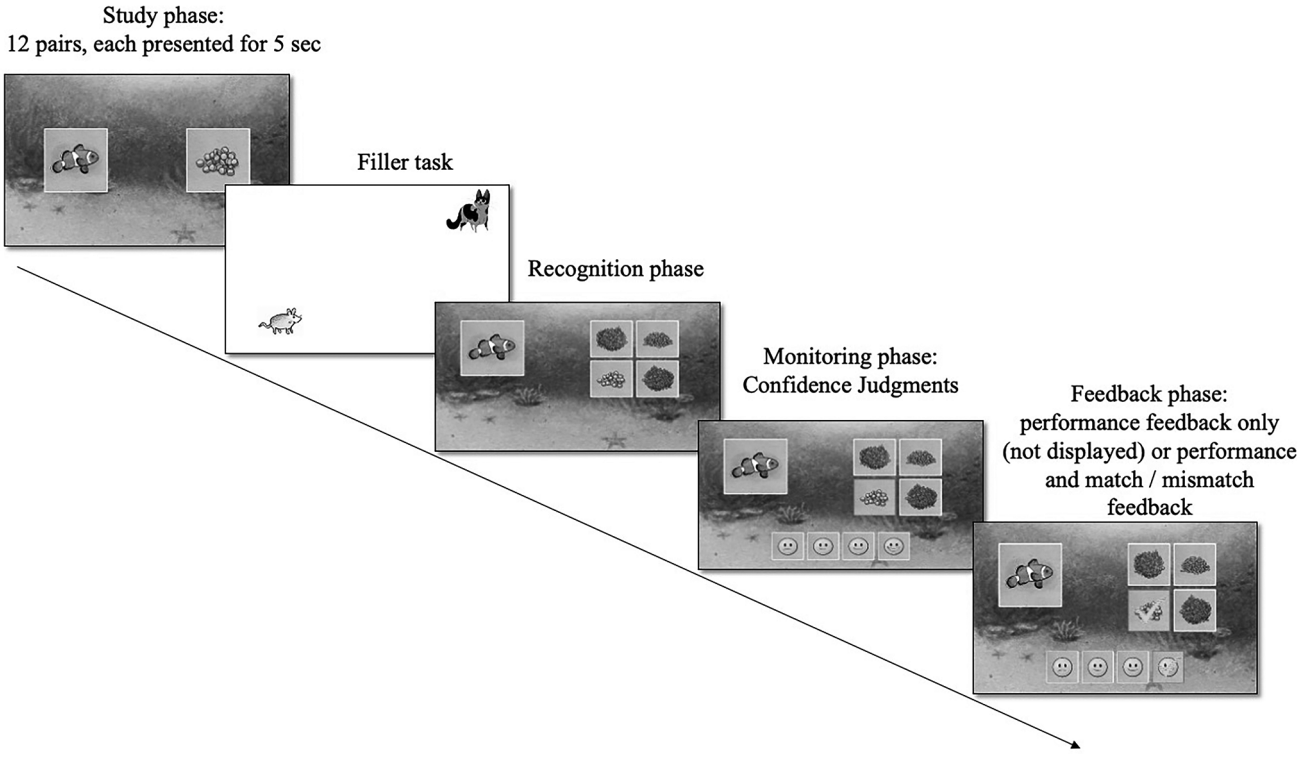

Figure 2. Procedure of the paired-associates task for consistency.

trated monitoring their recognition performance and the appropriate usage of the monitoring judgment scale. The participants received detailed explanations about which smiley face corresponded with each confidence level. The children easily understood the use of the scale. Then, the test phase with the 8 remaining items started. The participants selected one of the four foods for every item by touching the picture. A blue frame highlighted the chosen answer. If children wanted to keep their response, they were to tap a second time on their choice. Then the color of the frame changed from blue to yellow. However, if they wanted to change their choice, they could tap on another picture and tap again to confirm their selection until the frame color changed to yellow. The next item in each case was not presented to the children until they had selected an answer for the current item.

In the monitoring phase, immediately after every individual recognition trial, the children were asked for their confidence judgments by indicating how certain they were that their answer was correct. The accuracy of metacognitive judgments is often assessed with retrospective judgments of confidence (Butterfield & Metcalfe, 2001; Kelley & Lindsay, 1993). Using confidence judgments, learners indicate how confident they are that they have recalled the item correctly (Roebers, 2002). We used a 4-point smiley scale ranging from 1 = *very uncertain* to 4 = *very certain* to avoid a neutral judgment for which feedback would be very difficult to give.

In the feedback phase, immediately following the recognition, we provided one of two versions of visual feedback (green tick after correct response, red cross after incorrect response) and audio feedback (via headphones). We assigned every school class randomly to one of the two item-specific feedback conditions. The children in the Performance Feedback condition were informed whether their response was correct or incorrect *(You have chosen the right food./You have chosen the wrong food. Don't worry. It wasn't easy.)* The children in the Monitoring Feedback condition received the same performance feedback and were additionally informed whether their performance matched their confidence judgments. This resulted in eight different forms of favorable and child-appropriate verbal feedback. If actual performance and confidence judgments correspond: *(1) You have chosen the right food. Very good that you are very certain. (2) You have chosen the right food. Good that you are certain. (3) You have chosen the wrong food. Very good that you are very uncertain. (4) You have chosen the wrong food. Good that you are uncertain.* If actual performance and confidence judgments did not correspond: *(5) You have chosen the right food. It is a pity that you are very uncertain. (6) You have chosen the right food. It is a pity that you are uncertain. (7) You have chosen the wrong food. Don't worry. But it is a pity that you are very*

certain. (8) You have chosen the wrong food. Don't worry. But it is a pity that you are certain.

Measures of Monitoring Accuracy

To answer our main research questions, we focused on two monitoring measures: relative monitoring accuracy and absolute monitoring accuracy. While the first quantifies whether children's confidence judgments accurately predict item-specific performance (higher judgments for correct than for incorrect answers), the second focuses on the degree to which the level of confidence judgment corresponds to actual performance. We calculated a discrimination score for relative monitoring accuracy (Bol & Hacker, 2012; Nelson, 1996; Schraw, 2009). Separately for every session, we subtracted the mean of confidence judgments after correct responses from the mean of confidence judgments after incorrect responses. Positive discrimination scores indicate that the individual can judge whether the given answers were correct or incorrect. Negative discrimination scores indicate that the individual is poor in metacognitively discriminating between correct and incorrect answers. An increasing discrimination score over time would mirror a growing metacognitive awareness of correct versus incorrect responses.

Measures of absolute accuracy provide information on the individuals' judgments about the overall performance by comparing the confidence judgments and the actual performance of an item. It can be interpreted as whether an individual is overconfident, well-calibrated, or underconfident (Dunlosky & Metcalfe, 2009). The bias score is a common indicator of absolute monitoring accuracy. Although this is sometimes critically discussed in the literature (Hertzog et al., 2002), we decided to subtract the actual performance from the predicted confidence, yielding an unweighted individual bias score because of the limited number of items children could learn. Accordingly, the actual performance was coded as 0 % for incorrect responses and 100 % for correct responses. We recoded the confidence judgments as a percentage of postdicted recall; 1 (labeled *very uncertain*) = 0 %, 2 = 25 %, 3 = 75 %, 4 (labeled *very certain*) = 100 %). A positive value evidences overconfidence, whereas a negative value evidences underconfidence. Values around zero indicate perfect calibration (Dunlosky & Metcalfe, 2009; van Loon et al., 2013)

Performance

We assessed recognition performance as the percentage of correctly recognized items out of the 8 items per session. For the analyses reported below, recognition performance for each session served as a covariate to account for performance differences within feedback conditions and across sessions (Vuorre & Metcalfe, 2022).

Analysis Plan

We ran the analyses using the statistical software R version 4.2.1 (R Core Team, 2022). Given the nested structure of the data, i. e., repeated measures across 6 sessions (Level-1) which were nested within students (Level-2) and nested within classes (Level-3), we applied multilevel analysis as the primary statistical approach. First, we analyzed the effect of two different forms of feedback (performance feedback vs. monitoring feedback) on metacognitive discrimination and monitoring bias. Second, we addressed changes in confidence judgments after correct and incorrect answers, respectively, over time. We used the R package *lme4* version 1.1.30 with the function *lmer* (Bates et al., 2015) for both analyses. Multilevel models (MLM) can deal with violations of the assumption of independence of repeated observations. Furthermore, MLM allows all participants to be included in the analyses even when single data points are missing (Hox et al., 2017). To determine the children's relative monitoring accuracy, we employed the discrimination score as the dependent variable, session (measurement occasion) was the predictor variable at Level-1 (repeated measures), and the feedback condition (performance feedback vs. monitoring feedback) was the predictor at Level-2 (students' level). The Performance Feedback condition was defined as the reference category. We used bias as the dependent variable to examine absolute monitoring accuracy, that is, possible changes in overconfidence over time. The predictor variables at Level-1 and Level-2 were identical to the analysis of relative monitoring accuracy.

We also took an exploratory approach to acquire insights into the progress of monitoring abilities. Thus, we analyzed the effect of feedback on confidence judgments throughout the training separately for both feedback conditions and recognition correctness. As a dependent variable, we used mean confidence judgments derived from the smiley scale for correct and incorrect answers. We ran separate analyses for correct and incorrect responses, respectively, separately for the Performance Feedback condition and Monitoring Feedback condition, with mean confidence judgments as the dependent variable. Again, session was the predictor variable at Level-1. We conducted likelihood ratio tests for all MLM to compare the fit of nested models. Because of parsimony considerations, we only report the final MLMs.

Table 1. Means of performance, bias, discrimination, and confidence judgments by feedback condition and session

	Performance %	Bias	Discrimination	CJ correct responses	CJ incorrect responses
Performance Feedback					
T1	54.08 (0.19)	0.16 (0.28)	0.36 (0.68)	3.27 (0.61)	2.96 (0.74)
T2	44.50 (0.19)	0.21 (0.31)	0.43 (0.61)	3.20 (0.80)	2.78 (0.84)
T3	55.73 (0.18)	0.17 (0.32)	0.50 (0.76)	3.37 (0.65)	3.08 (0.85)
T4	37.76 (0.18)	0.31 (0.30)	0.11 (0.78)	3.14 (0.78)	2.97 (0.87)
T5	42.89 (0.17)	0.29 (0.30)	0.26 (0.74)	3.29 (0.80)	3.03 (0.75)
T6	47.16 (0.20)	0.27 (0.35)	0.42 (0.59)	3.45 (0.75)	3.05 (0.91)
Monitoring Feedback					
T1	67.36 (0.17)	0.08 (0.19)	0.32 (0.59)	3.48 (0.45)	3.06 (0.67)
T2	45.04 (0.17)	0.22 (0.20)	0.54 (0.62)	3.31 (0.56)	2.72 (0.61)
T3	65.28 (0.19)	0.07 (0.21)	0.58 (0.77)	3.46 (0.47)	2.79 (0.78)
T4	43.97 (0.22)	0.21 (0.24)	0.35 (0.77)	3.16 (0.59)	2.71 (0.63)
T5	57.97 (0.20)	0.12 (0.23)	0.42 (0.65)	3.32 (0.49)	2.88 (0.66)
T6	55.13 (0.21)	0.16 (0.25)	0.47 (0.89)	3.46 (0.49	2.95 (0.69)

Note. CJ = confidence judgments, *SD* in parentheses.

Results

Preliminary Analyses

Table 1 presents the percentage of correct recognition as a function of feedback conditions and training sessions. Performance across sessions varied between 37 % and 67 % correct answers, providing a well-suited database of confidence judgments for correct and incorrect answers throughout. Table 1 also displays monitoring accuracy (i. e., monitoring discrimination) and the mean confidence judgments after correct and incorrect responses, respectively, in each condition and session. As expected, we found a positive discrimination score for each session and in both feedback conditions, indicating that, on average, participants seemed to be able to discriminate between their correct and incorrect answers (relative monitoring accuracy: $t(111) = -14.69$, $p < .001$, $d = -0.89$). Furthermore, the bias scores for all sessions were positive, suggesting overall overconfidence (absolute monitoring accuracy; $t(111) = 9.40$, $p < .001$). Unexpectedly, descriptive statistics revealed that the performance in the Monitoring Feedback condition was higher than in the Performance Feedback condition. A two-way repeated measures ANOVA showed significant main effects for condition, $F(1,73) = 9.67$, $p = .01$, $\eta^2 = .04$, and session, $F(5,365) = 17.28$, $p < .001$ $\eta^2 = .14$. However, the interaction effect was not significant, $F(5,365) = 1.675$, $p = .14$. Pairwise comparisons (i. e., Bonferroni corrected *t*-tests), showed that performance was significantly higher in the Monitoring Feedback than in the Performance Feedback condition at session T_1 ($p < .001$), T_3 ($p = .01$), and T_5 ($p < .001$), but not at T_2 ($p = .88$), T_4 ($p = .12$), and T_6 ($p = .06$).

Monitoring Discrimination

To address our research question of whether children may benefit from item-specific feedback, we estimated a three-level, intercept-only model to calculate the intra-class correlation coefficient (ICC; repeated measures (Level-1), students (Level-2), and classes (Level-3). The ICC revealed that classes explain only 1 % of the intercept variance (ICC = .01). The model comparison revealed that the likelihood ratio test was not significant, $\chi^2(1) = 0.60$, $p = .44$. That indicates a two-level model has a better fit to the data than a three-level model. Because of the low number of classes and the lack of variance at Level-3, we used two-level models in the subsequent analyses. The results of all MLM conducted are presented in Table A1 (see Appendix A). For parsimony, we report only the random-coefficient model. The intercept-only model (Model 1) revealed significant between-student variation. The student level explained 10 % of the variance in the outcome variable (ICC = .10), that is, of all the variance observed, 10 % was attributable to individual differences between children. To analyze whether the feedback condition affected relative monitoring accuracy (i. e., discrimination) over six weeks, we tested a conditional model with Session as a predictor at Level-1 and Feedback condition as a predictor at Level-2, controlled for recog-

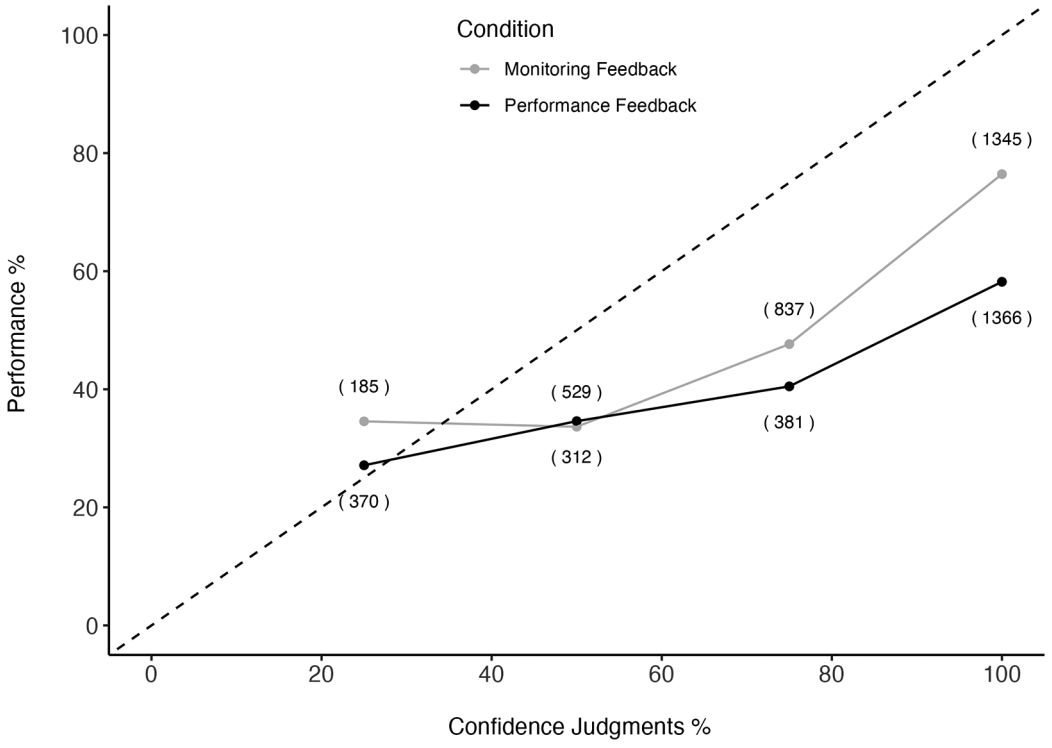

Figure 3. Calibration curve as a function of the feedback condition; frequencies in parentheses.

nition performance. Results showed that, against our expectations, the linear rate of change for discrimination was not significant, γ_{10} = –0.00, *SE* = 0.02, 95% CI [–0.04, 0.03], $t(529) = -.19$, $p = .85$. That is, the children's relative monitoring accuracy did not improve throughout the training. Further, the main effect of condition was not significant, γ_{11} = –0.06, *SE* = 0.07, 95% CI [–0.07, 0.19], $t(128) = 0.88$, $p = .38$. This indicates that there was no specific effect in either feedback condition on relative monitoring accuracy, which was also not in line with our expectations. According to the analysis results concerning relative monitoring accuracy, the discrimination scores remained constant over time. The fit of a cross-level interaction model was not significant.

Monitoring Bias

To test our hypothesis that children remain overconfident, we first inspected the data descriptively by plotting a calibration curve as a function of the feedback condition (Figure 3). Deviations below the dotted line (perfect calibration) indicate overconfidence, whereas deviations above the line indicate underconfidence.

As Figure 3 shows, on average, the children in both conditions were overconfident in judging their performance. Next, we ran MLMs for inferential statistics. Table B1 presents the results of all MLMs conducted (see Appendix B). We tested a slope-as-outcome model with cross-level interaction to examine whether feedback significantly affects the bias score over time. The simple slope of condition revealed that children in both conditions did not differ in their bias score at the beginning of the training, γ_{11} = –0.03, *SE* = 0.04, 95% CI [–0.11, 0.05], $t(112) = -0.73$, $p = .46$. The fit of a cross-level interaction model was significant. In the Monitoring Feedback condition, the results showed a linear session-by-session decrease in bias, which was about 0.02 lower than in the Performance Feedback condition, γ_{13} = - 0.02, *SE* = 0.04, 95% CI [–0.04, –0.01], $t(112) = -2.21$, $p = .03$. That means that, compared to the Performance Feedback condition, the children who received monitoring feedback significantly reduced their overconfidence over time. Nevertheless, the children in both conditions remained overconfident, indicated by a positive bias score, which aligned with our hypothesis.

Confidence Judgments

Considering that we did not find an effect of feedback on monitoring accuracy over time, we decided to explore the relationship between confidence judgments and recognition correctness (correct vs. incorrect responses) sep-

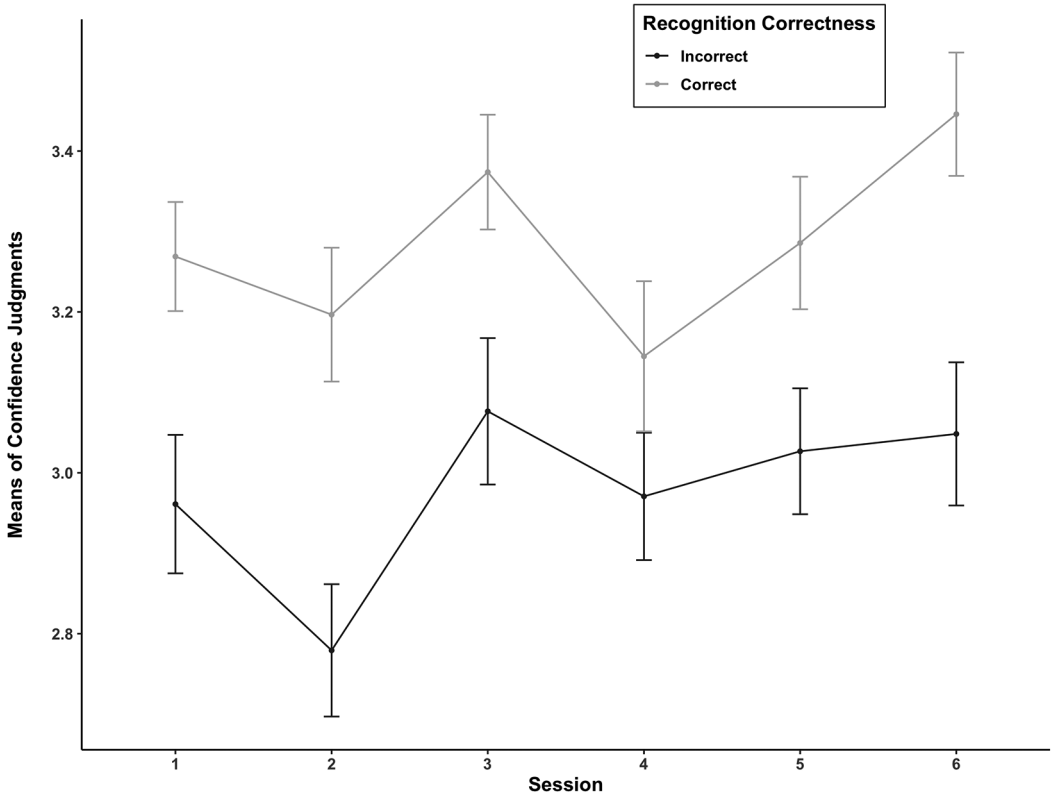

Figure 4. Changes in mean confidence judgments for performance feedback condition as a function of session.

arately for both feedback conditions. In other words, it may still be the case that (1) confidence judgments after correct and after incorrect responses change differently over time and (2) confidence judgments are affected differentially in the two feedback conditions, but the discrimination scores may mask this effect. Thus, further exploratory analyses seem worthwhile.

To explore the effect of the two qualitatively different forms of feedback on confidence judgments throughout the training, we conducted MLMs with session as the only predictor at Level-1 and Performance as time-variant Level-1 covariate. Table 1 shows the descriptive statistics of confidence judgments after correct and incorrect answers. Higher confidence judgments after correct and lower after incorrect responses would still indicate more fine-tuned monitoring skills. Thus, we separated the data by the participants' conditions and by recognition correctness.

Performance Feedback Condition

Before testing the unconditional predictor model, we compared the fit of an unconditional linear model to an unconditional quadratic model. The deviance test showed a better fit for the linear model, $\chi^2(1) = 1.96$, $p = .16$. Thus, we retained a linear model in all subsequent models. Table C1 represents the results of MLM for confidence judgments in the Performance Feedback condition after correct and incorrect responses (see Appendix C). The random-intercept model for *correct* recognition controlled for performance was significant, $\gamma_{10} = 0.03$, $SE = 0.02$, 95 % CI [0.00, 0.06], $t(1047) = 2.00$, $p = .05$. The positive slope indicated that, on average, the children's confidence for correct answers increased over time by 0.03 points. The fit of a random-intercept/random-slope model was not significant. The random-intercept/random-slope model for *incorrect* recognition controlled for performance also showed a significant increase in confidence judgments across the sessions, $\gamma_{10} = 0.05$, $SE = 0.02$, 95 % CI [0.00, 0.09], $t(52) = 2.08$, $p = .04$. The results indicate that, despite the children being made aware of their incorrect answers, their confidence judgments increased over time. Figure 4 shows the mean of confidence judgments for the Performance Feedback condition during 6 sessions for correct and incorrect responses.

Monitoring Feedback Condition

Table D1 represents the results of MLM for confidence judgments of the Monitoring Feedback condition after correct or incorrect responses (see Appendix D). Before testing the predictor model for correct and incorrect

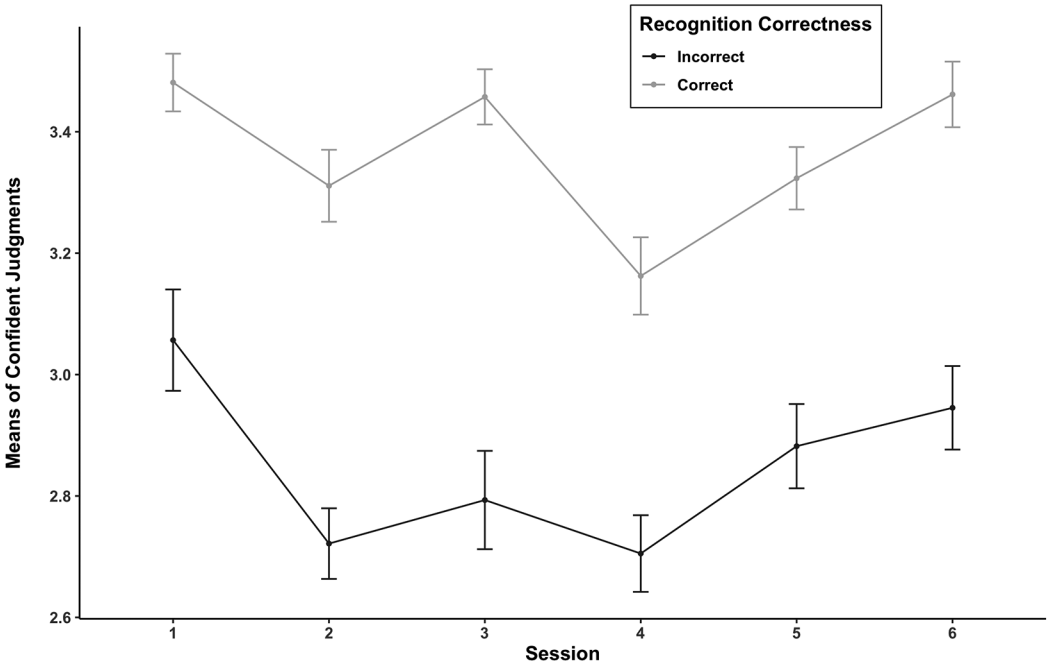

Figure 5. Changes in mean confidence judgments for monitoring feedback condition as a function of session.

responses, we compared the fit of linear models (Model 2) to quadratic models (Model 3). The deviance tests showed a better fit for the quadratic models. Thus, in all subsequent models, we retained quadratic models. The quadratic random-intercept/random-slope model (Model 3) for *correct* recognition controlled for performance showed a significant initial linear decrease (negative slope effect) $\gamma_{10} = -0.12$, $SE = 0.04$, $t(1458) = -2.86$, $p = .01$, 95% CI [–0.20, –0.04], followed by an accelerated increase (positive curve effects), $\gamma_{20} = 0.02$, $SE = 0.01$, 95% CI [0.01, 0.04], $t(1454) = 2.73$, $p = .01$, indicating that, between session 1 to session 2, on average, confidence judgments after correct responses tended to decrease by 0.12 points. The quadratic random-intercept/random-slope model (Model 3) for *incorrect* response controlled for performance showed a significant initial linear decrease (negative slope effect), $\gamma_{10} = -0.17$, $SE = 0.06$, $t(900) = -2.88$, $p < .01$, with variability in the linear rate of change (95% CI [–0.29, –0.05], followed by an accelerated decline (positive curve effects), $\gamma_{20} = 0.03$, $SE = 0.01$, 95% CI [0.01, 0.06], $t(1155) = 3.22$, $p < .001$, indicating that, between session 1 to session 2, on average, confidence judgments after incorrect responses tended to decrease by 0.17 points. However, the decrease at the beginning was not long-lasting because the linear effect was not constant and the quadratic coefficient was positive. Confidence judgments tended to increase again after half of the training. Figure 5 shows the mean of confidence judgments for the Monitoring Feedback condition during 6 sessions for correct and incorrect responses.

Discussion

We designed the current training study to improve metacognitive monitoring judgments by providing two different types of feedback. The 7- to 8-year-olds solved paired-associates memory-learning task over 6 sessions (8 items per task) and rated their confidence about the correctness of each response. In one condition, the children received feedback about their performance (performance feedback). In the other condition, the children received feedback on their performance *and* on whether their performance matched their monitoring judgment (monitoring feedback). Thus, the children made numerous metacognitive experiences to connect their performance with their monitoring accuracy. This contribution focuses on possible changes in monitoring accuracy over the course of the 6 sessions by including both relative and absolute monitoring accuracy. In line with our expectations, the results of relative monitoring accuracy revealed a positive discrimination score, which held true for each session. In other words, the children in both conditions reported higher confidence when their response was correct and lower confidence when their response was incorrect. Our results align with previous research demonstrating that even preschool-aged children can metacognitively discriminate between correct and incorrect answers in their monitoring judgments (Gonzales et al., 2022; Lyons & Ghetti, 2013).

We expected that multiple sessions with child-appropriate feedback on either performance or performance

and monitoring would – sooner or later – yield to better metacognitive discrimination. However, our results revealed that the children in both feedback conditions did not improve their monitoring discrimination throughout the training. This finding contrasts with research emphasizing the role of metacognitive experiences (Efklides, 2011). Considering the few existing studies in which, through face-to-face feedback, children learned to recognize more errors (Destan & Roebers, 2015; Lipko-Speed et al., 2018), it appears that even the children in the Monitoring Feedback condition could not implement our feedback to develop more fine-tuned adjustments toward better-calibrated monitoring processes. As expected, the children's bias scores lay uniformly beneath the optimal calibration line indicating overconfidence (see Figure 3). Even though the children in the Monitoring Feedback condition significantly decreased their overconfidence over time, the average degree of certainty nevertheless pointed toward overconfidence (poor absolute monitoring accuracy). This underscores the necessity of differentiating between absolute and relative monitoring accuracy. The present study once again shows that, although relative monitoring accuracy was observable, absolute monitoring accuracy was poor (Schraw, 2009). From an evolutionary perspective, overconfidence might be adaptive for a child (Bjorklund & Bering, 2002; Shin et al., 2007). According to the empirical literature on wishful thinking, children predict future performance based on desired rather than realistic outcomes. (e.g., Schneider, 1998; Serra & DeMarree, 2016). In any educational context, however, item-specific overconfidence has repeatedly been shown to be detrimental for detecting and correcting errors (Bayard et al., 2021; Destan & Roebers, 2015), as well as for efficient self-regulated learning in general (Efklides, 2006). Children's overconfidence in this and many other studies makes clear that training to reduce overconfidence is needed.

We then further explored whether the children in either condition might have become less certain overall and thus analyzed confidence for correct and incorrect responses separately. In the Performance Feedback condition, there was a shift in confidence toward even higher confidence, both for correct and incorrect responses, possibly suggesting motivational aspects such as self-protection in the face of repeated (mild) negative feedback without direct negative consequences (Bjorklund & Green, 1992). However, it may be beneficial when overconfidence motivates children to continue with the task. As well as promoting some self-protection, it can also help children to develop a sense of mastery and accomplishment (Händel & Bukowski, 2019; Händel & Fritzsche, 2016; Schneider & Lockl, 2008). In the Monitoring Feedback condition, in contrast, we detected a quadratic term for the children's monitoring of correct and incorrect responses, suggesting that, during the first sessions, the children became less confident. However, over the course of the training, paradoxically, the participants returned to their initial overconfidence. With this shift toward higher confidence in the second half of the training, it becomes difficult to separate the effects of performance and monitoring feedback, although qualitatively, the two feedback conditions differ substantially. Future research should examine which aspects of feedback may cause a shift from uncertainty to certainty and vice versa.

Why did children *not* benefit from the metacognitive experiences, including detailed feedback on their monitoring accuracy? For one, although young children receive a large amount of feedback in their natural environment, for example, whether or not they are meeting certain standards of their caregivers, formal learning in traditional classrooms seldom provides item-specific, detailed feedback on performance and only very rarely, if ever, on monitoring accuracy. Moreover, and as Efklides (2011) emphasizes, the social aspect of feedback should not be underestimated. In our case, feedback was provided via screen and headphones – personalized but still automated – and this missing social interaction might have contributed to the lack of a training effect. Furthermore, receiving feedback not from a social partner but rather from the computer may have been very unusual for the participants, and they might have had difficulties in deeply processing this kind of feedback (Butler & Winne, 1995). Yet, this interpretation seems unlikely since the children easily mastered the task and very seldom (actually never after the first session) asked for help. It also should be noted that the learning task was designed for research purposes only and was not part of the children's obligatory school curriculum. Children may assess themselves differently and interpret feedback distinctly through interaction with peers and teachers. Future research should address metacognitive judgments during actual classroom learning tasks (e.g., math tasks) and examine these effects on school achievement.

Furthermore, the lack of monitoring accuracy improvements may possibly be attributed to the answer format. Greving and Richter (2018), for example, demonstrated that the presence of distractors in multiple-choice tasks, in contrast to short-answer questions, negatively affected memory retention and hampered university students' learning. However, providing item-specific corrective feedback to second-graders could diminish the negative effects of multiple-choice testing (Marsh et al., 2012). Notwithstanding possible negative effects on test performance, multiple-choice test formats are popular performance assessments (Butler, 2018). For future research on the impact of feedback, multiple-choice tests should be

carefully designed by integrating both feedback that provides the correct solution and feedback on metacognitive judgments. Another reason might lie in the number of training sessions: The number of sessions with feedback may still have been too low (for example, compared to the executive function training literature where 20–30 sessions are the rule; Diamond & Ling, 2020). In the present study, the children received feedback after each response in the memory test (8 items each in 6 sessions), and yet the children did not benefit. They received far more feedback and item-specific feedback than the participants in previous studies, and previous studies gave only global feedback on performance pre- and postdictions (Lipko et al., 2009; Lipko-Speed et al., 2018; van Loon & Roebers, 2020). Altogether, these and other results indicate that children are relatively resistant to corrections of their overconfidence, probably independently of how feedback is transmitted.

Given the sample size, it was impossible to address individual differences in how the feedback affected monitoring accuracy. Van Loon and Roebers (2020) found that working memory might be such a moderating variable, influencing how well a child processes the provided feedback. We sought to counteract such an effect by providing visual and audio feedback, but working memory and potentially other individual differences might affect feedback effectiveness differentially. Thus, we cannot exclude the possibility that the training might yield differential effects. Future studies should include individual differences in cognitive abilities, working memory, and possibly also instruction comprehension. Another explanation for the missing training effect may lie in children's reliance on invalid or their nonreliance on valid mnemonic cues. The cue-utilization framework assumes that learners rely on internal, subjective indicators (mnemonic cues) when monitoring their performance (Koriat, 1997). In recent work, there is increasing evidence that even young children have such mnemonic cues at their disposal and use them when monitoring (Bayard et al., 2021; Geurten et al., 2018, 2021). Thus, an interesting question arises: Would training be more effective if the feedback also addressed task-inherent cues (e.g., "It took you a long time to select one of the alternatives for this item, so are you very certain?")? In a review regarding feedback effects in an educational context, Shute (2008) found that the level of information that aims to modify confidence should address performance accuracy and information that allows learners to discover strategies and, in our case, possibly mnemonic cues. Adapting the feedback may constitute another avenue for further training research in the present domain.

This was the first attempt at computer-based training targeting young children's monitoring accuracy. Of course, some limitations need to be considered when discussing our findings. For one, the monitoring discrimination score was calculated based on a relatively small number of items, so that calculation of gamma correlations was impossible. Moreover, participants in the two feedback conditions differed significantly in terms of their performance on the paired-associates task for consistency (possibly because of class-wise group assignments). Performance differences are known to impact monitoring accuracy (Vuorre & Metcalfe, 2022). Differences in more general cognitive abilities may also have been driving the memory differences as well as influencing monitoring accuracy (van der Stel & Veenman, 2014). In line with this, prior studies found that monitoring abilities and intelligence are associated, and thus, intelligence might have been a moderating factor here (Ohtani & Hisasaka, 2018). Another issue not addressed in this study is that sessions were not administered in a counterbalanced order, and some sessions were more difficult than others in terms of overall difficulty. Yet, posthoc item-difficulty analyses revealed that, within every session, there were items with varying degrees of difficulty, ensuring that, within every session, differences in item-retrieval fluency were at the children's disposal. Lastly, our study included two qualitatively different experimental conditions but not a control condition. However, given that children uniformly show poor relative and absolute monitoring accuracy, only a wait-list control condition would have been ethically justifiable.

This is one of the very first training studies aiming to improve relative monitoring accuracy in primary school children. The computer-based application has a clear theoretical background, making such training feasible for classroom implementation and easy to adapt in future studies. The results extend the existing research by showing how children can make many memory experiences and exercise metacognitive monitoring in a child-appropriate way. Children's struggle with correcting their monitoring and their resistance to taking the feedback into account was surprising. If we zoom into the monitoring of correct and incorrect responses, our results still seem to suggest that feedback can bias children's monitoring judgments positively and negatively. We conclude that monitoring feedback – as long as it was new to the children – seemed to trigger advanced information-processing abilities that may lead to more sophisticated monitoring skills in the long run (Kluger & DeNisi, 1996). Repeated feedback about children's performance seems to unintentionally increase children's overconfidence, perhaps because of implicit self-protection strategies.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1 – 48. https://doi.org/10.18637/jss.v067.i01

Bayard, N. S., van Loon, M. H., Steiner, M., & Roebers, C. M. (2021). Developmental improvements and persisting difficulties in children's metacognitive monitoring and control skills: Cross-sectional and longitudinal perspectives. *Child Development*, *92*(3), 1118 – 1136. https://doi.org/10.1111/cdev.13486

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist, 47*, 46 – 54. https://doi.org/10.1037/0003-066X.47.1.46

Bjorklund, D. F., & Bering, J. M. (2002). The evolved child: Applying evolutionary developmental psychology to modern schooling. *Learning and Individual Differences*, *12*(4), 347 – 373. https://doi.org/10.1016/S1041-6080(02)00047-X

Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebers, C. M. (2021). Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning, 16*(3), 749 – 768. https://doi.org/10.1007/s11409-021-09261-z

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323 – 331. https://doi.org/10.1016/j.jarmac.2018.07.002

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

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1491 – 1494. https://doi.org/10.1037/0278-7393.27.6.1491

Bol, L., & Hacker, D. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 1 – 6. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00229

Dapp, L. C., & Roebers, C. M. (2021). Metacognition and self-concept: Elaborating on a construct relation in first-grade children. *PLOS ONE*, *16*(4), e0250845. https://doi.org/10.1371/journal.pone.0250845

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347 – 374. https://doi.org/10.1007/s11409-014-9133-z

Destan, N., Spiess, M. A., de Bruin, A., van Loon, M., & Roebers, C. M. (2017). 6- and 8-year-olds' performance evaluations: Do they differ between self and unknown others? *Metacognition and Learning, 12*(3), 315 – 336. https://doi.org/10.1007/s11409-017-9170-5

Diamond, A., & Ling, D. S. (2020). Review of the evidence on, and fundamental questions about, efforts to improve executive functions, including working memory. In J. M. Novick, M. F. Bunting, M. R. Dougherty, & R. W. Engle (Eds.), *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development* (pp. 143 – 431). Oxford University Press. https://doi.org/10.1093/oso/9780199974467.003.0008

Dignath, C., Buettner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively?: A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*(2), 101 – 129. https://doi.org/10.1016/j.edurev.2008.02.003

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228 – 232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

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

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271 – 280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Efklides, A. (2006). Metacognitive experiences: The missing link in the self-regulated learning process. *Educational Psychology Review*, *18*(3), 287 – 291. https://doi.org/10.1007/s10648-006-9021-4

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277 – 287. https://doi.org/10.1027/1016-9040.13.4.277

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6 – 25. https://doi.org/10.1080/00461520.2011.538645

Efklides, A., & Metallidou, P. (2020). Applying metacognition and self-regulated learning in the classroom. In *Oxford Research Encyclopedia of Education.* Retrieved from https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-961

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

Geurten, M., Lloyd, M., & Willems, S. (2017). Hearing "quack" and remembering a duck: Evidence for fluency attribution in young children. *Child Development*, *88*(2), 514 – 522. https://doi.org/10.1111/cdev.12614

Geurten, M., Meulemans, T., & Willems, S. (2018). A closer look at children's metacognitive skills: The case of the distinctiveness heuristic. *Journal of Experimental Child Psychology*, *172*, 130 – 148. https://doi.org/10.1016/j.jecp.2018.03.007

Geurten, M., Willems, S., & Lloyd, M. (2021). Too much familiarity! The developmental path of the fluency heuristic in children. *Child Development*, *92*(3), 919 – 936. https://doi.org/10.1111/cdev.13449

Gonzales, C. R., Merculief, A., McClelland, M. M., & Ghetti, S. (2022). The development of uncertainty monitoring during kindergarten: Change and longitudinal relations with executive function and vocabulary in children from low-income backgrounds. *Child Development*, *93*(2), 524 – 539. https://doi.org/10.1111/cdev.13714

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

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., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition*, *44*(2), 229 – 241. https://doi.org/10.3758/s13421-015-0552-0

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

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging*, *17*(2), 209 – 225. https://doi.org/10.1037/0882-7974.17.2.209

Hox, J., Moerbeek, M., & Schoot, R. van de. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

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*(4 – 5), 689 – 717. https://doi.org/10.1080/09541440701326170

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*(1), 1 – 24. https://doi.org/10.1006/jmla.1993.1001

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254 – 284. https://doi.org/10.1037/0033-2909.119.2.254

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349 – 370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development*, *24*(2), 169 – 182. https://doi.org/10.1016/j.cogdev.2009.01.001

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 34 – 53. https://doi.org/10.1037/0278-7393.27.1.34

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152 – 166. https://doi.org/10.1016/j.jecp.2008.10.002

Lipko-Speed, A. R., Buchert, S., & Merriman, W. E. (2018). Observing an adult model can cause immediate improvement in preschoolers' knowledge judgments. *Cognitive Development*, *48*, 225 – 234. https://doi.org/10.1016/j.cogdev.2018.09.003

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development*, *84*(2), 726 – 736. https://doi.org/10.1111/cdev.12004

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*(8), 899 – 906. https://doi.org/10.1080/09658211.2012.708757

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

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1 – 12. https://doi.org/10.1016/j.jml.2013.09.007

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125 – 173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item comments on Schraw (1995). *Applied Cognitive Psychology, 10*(3), 257 – 260. https://doi.org/10.1002/(SICI)1099-0720(199606)10:3<257::AID-ACP400>3.0.CO;2-9

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*(2), 159 – 179. https://doi.org/10.1007/s10409-006-9595-6

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179 – 212. https://doi.org/10.1007/s11409-018-9183-8

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The Journal of Experimental Education*, *65*(3), 213 – 228. https://doi.org/10.1080/00220973.1997.9943455

R Core Team. (2022). *A language and environment for statistical computing. R foundation for statistical computing* (Version 4.2.1) [Computer software]. Retrieved from https://www.R-project.org/

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning*, *5*(3), 229 – 250. https://doi.org/10.1007/s11409-010-9059-z

Roebers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38*(6), 1052 – 1067. https://doi.org/10.1037/0012-1649.38.6.1052

Roebers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental Psychology*, *55*(10), 2077 – 2089. https://doi.org/10.1037/dev0000776

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science*, *1*(2), 291 – 297. https://doi.org/10.1111/1467-7687.00044

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence from developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (Vol. 14, pp. 391 – 409). Taylor & Francis.

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

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*(7), 1127 – 1137. https://doi.org/10.3758/s13421-016-0624-9

Shin, H., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development*, *22*(2), 197 – 212. https://doi.org/10.1016/j.cogdev.2006.10.001

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153 – 189. https://doi.org/10.3102/0034654307313795

Tsalas, N., Paulus, M., & Sodian, B. (2015). Developmental changes and the effect of self-generated feedback in metacognitive controlled spacing strategies in 7-year-olds, 10-year-olds, and adults. *Journal of Experimental Child Psychology*, *132*, 140 – 154. https://doi.org/10.1016/j.jecp.2015.01.008

van der Stel, M., & Veenman, M. V. J. (2014). Metacognitive skills and intellectual ability of young adolescents: A longitudinal study from a developmental perspective. *European Journal of Psychology of Education*, *29*(1), 117 – 137. https://doi.org/10.1007/s10212-013-0190-5

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects

primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, *24*, 15 – 25. https://doi.org/10.1016/j.learninstruc.2012.08.005

van Loon, M. H., & Roebers, C. M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. *Early Childhood Research Quarterly*, *53*, 301 – 313. https://doi.org/10.1016/j.ecresq.2020.05.007

van Loon, M. H., & Roebers, C. M. (2021). Using feedback to support children when monitoring and controlling their learning. In D. Moraitou & P. Metallidou (Eds.), *Trends and prospects in metacognition research across the life span* (pp. 161 – 184). Springer International Publishing. https://doi.org/10.1007/978-3-030-51673-4_8

Vuorre, M., & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*, *17*(2), 269 – 291. https://doi.org/10.1007/s11409-020-09257-1

History
Published online June 15, 2023

Acknowledgments
We want to thank the involved master's students and research assistants for their outstanding help with data collection. We would also like to thank Stefan Kodzhabashev for programming the tasks.

Conflict of Interest
The authors declare that they have no conflict of interest.

Publication Ethics
Ethical approval for the study was obtained from the Faculty's Ethics Committee (Faculty of Human Sciences, University of Bern; Approval No. 2020 – 10 – 00005).

Authorship
Florian J. Buehler and Claudia M. Roebers contributed to the study's conception and design. Florian J.Buehler mainly performed the material preparation and data collection. Kristin Kolloff performed the data analysis. Kristin Kolloff wrote the first draft of the manuscript, and the second author commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Open Data
The pre-registration for the main study can be found: https://osf.io/mwnsy.

ORCID
Kristin Kolloff
https://orcid.org/0000-0002-7472-0268
Claudia M. Roebers
https://orcid.org/0000-0003-3048-7609
Florian J. Buehler
https://orcid.org/0000-0003-3107-2042

Kristin Kolloff
Institute of Psychology
Faculty of Human Sciences
University of Bern
Fabrikstrasse 8
3012 Bern
Switzerland
kristin.kolloff@unibe.ch

Appendix A

Table A1. Model parameter and goodness of fit for linear changes in monitoring discrimination

Effect (parameter)	Model 1	Model 2	Model 3
	Fixed effects		
Intercept (γ_{00})	0.40*** (0.04)	0.41*** (0.09)	0.19*(0.10)
Session (γ_{10})		−0.01 (0.02)	−0.00 (0.02)
Feedback condition (γ_{11})			0.06 (0.07)
Recognition performance (γ_{12})			0.35* (0.14)
	Random effects		
Variance components			
Residual (σ_2)	0.46	0.46	0.44
Intercept (τ_{00})	0.05	0.05	0.01
Slope (τ_{11})			0.00
Covariance (ρ_{01})			1.00
ICC	0.10		
Marginal R^2/Conditional R^2		0.00/0.10	0.01/ 0.11
	Goodness of fit		
Deviance	1,339.80	1,339.5	1,326.40
$\Delta\chi^2$		0.30	13.10*
Δdf		1	4

Note. Standard errors in parentheses. All *p* values in this table are two-tailed. In Model 1 (intercept-only model), the intercept parameter estimate (γ_{00}) represents the average discrimination score across sessions. In Model 2 (random-intercept model), the intercept parameter estimate (γ_{00}) represents the average discrimination score at session 1 across students, γ_{10} represents the difference in discrimination score from a one unit increase in session (estimated rate of change). In Model 3 (random-coefficient model), the intercept parameter estimate (γ_{00}) represents the average discrimination score for the Performance Feedback condition at session 1, γ_{10} represents a linear rate of change from a one-unit increase in session for participants in the Performance Feedback condition, γ_{11}

Appendix B

Table B1. Model parameter and goodness of fit for linear changes in bias

Effect (parameter)	Model 1	Model 2	Model 3
	Fixed effects		
Intercept (γ_{00})	0.18*** (0.02)	0.24***(0.03)	0.21*** (0.03)
Session (γ_{10})		0.01** (0.01)	0.03*** (0.01)
Feedback condition (γ_{11})		−0.08† (0.04)	−0.03 (0.04)
Recognition performance (γ_{12})		−0.09*** (0.01)	−0.09*** (0.01)
Session * feedback condition (γ_{13})			−0.03* (0.01)
	Random effects		
Variance components			
Residual (σ_2)	0.03	0.02	0.02
Intercept (τ_{00})	0.04	0.05	0.05
Slope (τ_{11})		0.00	0.00
Covariance (ρ_{01})		−0.42	−0.41
ICC	0.57		
Marginal R^2/Conditional R^2		0.06/0.71	0.08/0.71
	Goodness of fit		
Deviance	-2,778.60	-4,698.40	-4,703.20
$\Delta\chi^2$		1,919.90***	495.78***
Δdf		3	3

Note. Standard errors in parentheses. All *p*-values in this table are two-tailed. In Model 1 (intercept-only model), the intercept parameter estimate (γ_{00}) represents the average bias score across sessions. In Model 2 (random-coefficients model), the intercept parameter estimate (γ_{00}) represents the average bias score at session 1 across students, γ_{10} represents the difference in bias score from a one unit increase in session (estimated rate of change), γ_{11} represents the difference in bias score between the Performance Feedback condition and the Monitoring Feedback condition by one unit increase of time. In Model 3 (slope-as-outcome), the intercept parameter estimate (γ_{00}) represents the average bias score for the Performance Feedback condition at session 1, γ_{10} represents the linear rate of change from a one unit increase in time for participants in the Performance Feedback condition, γ_{11} represents the difference in bias score between the Performance Feedback condition and the Monitoring Feedback condition by one unit increase of time, γ_{13} represents the cross-level interaction of Condition by Time. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Appendix C

Table C1. Model parameter and goodness of fit for linear changes in CJ for the performance feedback condition

Effect (parameter)	CJ after correct responses			CJ after incorrect responses		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
	Fixed effects					
Intercept (γ_{00})	3.31*** (0.08)	3.01*** (0.14)	3.00***(0.10)	2.94*** (0.10)	2.74*** (0.14)	2.70*** (0.11)
Session (γ_{10})		0.03* (0.02)	0.03† (0.02)		0.05** (0.02)	0.05* (0.02)
Recognition performance (γ_{20})		0.42* (0.18)	0.44*(0.18)		0.20 (0.18)	0.27 (0.18)
	Random effects					
Variance components						
Residual (σ_2)	0.80	0.80	0.79	1.00	0.99	0.96
Intercept (τ_{00})	0.29	0.30	0.26	0.44	0.45	0.47
Slope (τ_{11})			0.00			0.01
Covariance (ρ_{01})			0.10			–0.24
ICC	0.27			0.31		
Marginal R^2/Conditional R^2		0.01/0.28	0.01/0.29		0.01/0.32	0.01/0.34
	Goodness of fit					
Deviance	2,944.30	2,936.30	2,933.60	3,585.30	3,577.40	3565.30
$\Delta\chi^2$		8.01*	2.71		7.92*	12.08**
Δdf		2	2		2	2

Note. Standard errors in parentheses. CJ = confidence judgments. All *p*-values in this table are two-tailed. In Model 1 (intercept-only model), the intercept parameter estimate (γ_{00}) represents the average CJ score after correct or incorrect responses across sessions. In Model 2 (random-intercept model), the intercept parameter estimate (γ_{00}) represents the average CJ score at session 1 across students, γ_{10} represents the average linear rate of change in CJ after correct or incorrect responses at session 1. The intercept was allowed to vary. In Model 3 (random-intercept/random-slope model), the intercept parameter estimate (γ_{00}) represents the average CJ score at session 1 across students, γ_{10} represents the average linear rate of change in CJ after correct or incorrect responses at session 1. The intercept and linear slope were allowed to vary. $^{\dagger}p \leq .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Appendix D

Table D1. Model parameter and goodness of fit for linear changes in CJ for the monitoring feedback condition

Effect (parameter)	CJ after correct responses			CJ after incorrect responses		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
	Fixed effects					
Intercept (γ_{00})	3.38*** (0.05)	2.99***(0.05)	3.08*** (0.06)	2.85***(0.06)	2.65***(0.11)	2.84*** (0.13)
Session (γ_{10})		−0.01 (0.01)	−0.12** (0.04)		0.01 (0.02)	−0.17** (0.06)
Session^2 (γ_{20})			0.02** (0.01)			0.03** (0.01)
Recognition performance (γ_{30})		0.66*** (0.12)	0.62*** (0.12)		0.35* (0.14)	0.24 (0.15)
	Random effects					
Variance components						
Residual (σ_2)	0.60	0.59	0.58	0.74	0.70	0.70
Intercept (τ_{00})	0.10	0.10	0.10	0.21	0.29	0.29
Slope (τ_{11})		0.00	0.00		0.01	0.01
Covariance (ρ_{01})		−0.14	−0.18		−0.51	−0.51
ICC	0.15			0.23		
Marginal R^2/Conditional R^2		0.02/0.17	0.03/0.17		0.01/0.26	0.01/0.27
	Goodness of fit					
Deviance	3,584.20	3,551.40	3,543.90	3,138.10	3,118.50	3108.20
$\Delta\chi^2$		32.81***	7.49**		19.55***	10.33**
Δdf		4	1		4	1

Note. Standard errors in parentheses. CJ = confidence judgments. All *p* values in this table are two-tailed. In Model 1 (intercept-only model), the intercept parameter estimate (γ_{00}) represents the average CJ score after correct or incorrect responses across sessions. In Model 2 (random-intercept model), the intercept parameter estimate (γ_{00}) represents the average CJ score after correct or incorrect responses at session 1 across students, γ_{10} represents the average linear rate of change in CJ after correct or incorrect responses at session 1. The intercept and slope were allowed to vary. In Model 3 (quadratic random-intercept/random-slope model), the intercept parameter estimate (γ_{00}) represents the average CJ score after correct or incorrect responses at session 1 across students, γ_{10} represents the average linear rate of change in CJ after correct or incorrect responses at session 1 from a one-unit change in session, (γ_{20}) represents acceleration–deceleration in each growth trajectory for every one-unit increase in CJ after correct or incorrect responses. The intercept, linear slope (session), and quadratic slope (session^2) were allowed to vary. *p < .05, **p < .01, ***p < .001.