ZDM (2019) 51:587–600
https://doi.org/10.1007/s11858-019-01062-8

ORIGINAL ARTICLE

Metacognition in mathematics: do different metacognitive monitoring measures make a difference?

Klaus Lingel[1] · Jan Lenhart[1] · Wolfgang Schneider[1]

Accepted: 10 May 2019 / Published online: 24 May 2019

Abstract
Metacognitive monitoring in educational contexts is typically measured by calibration indicators, which are based on the correspondence between cognitive performance and metacognitive confidence judgment. Despite this common rationale, a variety of alternative methods are used in the field of monitoring research to assess performance and judgment data and to calculate calibration indicators from them. However, the impact of these methodological differences on the partly incongruent picture of monitoring research has hardly been considered. Thus, the goal of the present study is to examine the effects of methodological choices in the context of mathematics education. To do so, the study compares the effects of two judgment scales (Likert scale vs. visual analogue scale), two response formats (open-ended response vs. closed response format), the information base of judgment (prospective vs. retrospective), and students' achievement level on confidence judgments. Secondly, the study contrasts measures of three calibration constructs, namely absolute accuracy (Absolute Accuracy Index, Hamann Coefficient), relative accuracy (Gamma, d'), and diagnostic accuracy (sensitivity and specificity). One hundred and nine seventh-grade students completed a set of 20 mathematical problems and rated their confidence in a correct solution for each problem prospectively and retrospectively. Our results show a pervasive overconfidence of students across achievement levels. Monitoring was more precise for retrospective judgments and the visual analogue scale format. Gamma, sensitivity, and specificity proved to be susceptible for boundary values, caused by the general overconfidence in the sample. Measures of absolute accuracy were affected by response format of the task and judgment scale, with higher accuracy found for closed response format and visual analogue scale. We observed substantial correlations within the three calibration constructs and comparably low correlations between indicators of different constructs, confirming three interrelated aspects of monitoring accuracy. The low correlations between corresponding prospective and retrospective calibration indicators suggest different calibration processes. Implications for studies on calibration and mathematics education are discussed.

Keywords Metacognition · Metacognitive monitoring · Calibration · Mathematics

1 Introduction

Research on metacognition originated in the domain of memory development (termed metamemory) and is theoretically, methodologically, and empirically well elaborated in this domain (Dunlosky and Tauber 2016). Its potential was also recognized early in other domains such as text comprehension (for a review, see Baker 1989) and mathematics (for a review, see Schneider and Artelt 2010), showing that students' metacognitive knowledge and competencies were substantially related to their performance.

Metacognition is defined as any knowledge or cognitive activity that takes cognitive processes as its object (cf. Flavell et al. 2002). Thus, on the one hand, metacognition refers to people's knowledge about their own information processing skills, about the nature of cognitive tasks, and about strategies for coping with such tasks. On the other hand, it also includes executive skills related to monitoring and self-regulation of one's own cognitive activities. With regard to mathematics instruction, the role of monitoring was especially emphasized (e.g., Desoete and Veenman 2006; Schoenfeld 1987).

As in other domains, an extensive repertoire of assessment methods has been developed in the field of mathematical

✉ Klaus Lingel
lingel@uni-wuerzburg.de

[1] Department of Psychology, University of Würzburg, Würzburg, Germany

metacognition (e.g., Desoete 2008). However, unlike the metamemory domain with its long-lasting and vivid discussions on methodological issues and problems (e.g., Schwartz and Metcalfe 1994; Dunlosky et al. 2016), little systematic research on the characteristics and possible shortcomings of different measures of metacognitive monitoring has been conducted in the domain of mathematics. Consequently, this study aims to examine the effects of methodological issues on monitoring assessment in mathematics education.

1.1 Monitoring in the domain of mathematics

In mathematics, Garofalo and Lester (1985) provided a seminal conceptualization of metacognitive monitoring during mathematical problem solving. Integrating ideas of Polya (1949) and Schoenfeld (1985), they differentiated four phases: (1) orientation assessing and understanding a problem in the orientation phase, (2) planning of solution behavior and choice of actions during the organization phase, (3) regulation of solution behavior during the execution phase, (4) evaluation of planning decisions and outcomes during the verification phase. Monitoring occurs in all four phases and refers to an ongoing evaluation of one's own cognitive activities, with the goal of initiating regulation processes (Schoenfeld 1985).

Typically, monitoring is assessed prospectively in the orientation phase, immediately preceding the execution of a cognitive task, or retrospectively in the verification phase, immediately following the execution of a cognitive task. Prospective judgments require an activation of knowledge about the task, about one's own abilities, as well as about adequate strategies and enable the individual to adapt time and effort. Retrospective judgments in mathematics require reflections on processes and outcomes, or more specifically, self-assessments of task understanding, of appropriateness of planning, executing and regulating the solution process (Garofalo and Lester 1985).

Thus, monitoring is a critical activity in mathematical problem solving. Erroneous monitoring, regardless of whether the judgments are over- or underconfident, may lead to deficiencies in the activation of relevant content knowledge and the regulation of cognitive processes (Hacker et al. 2008). As a consequence, the quality of monitoring affects in the short term the performance in the task at hand, and in the long term the accumulation of cognitive and metacognitive knowledge on mathematical problem solving.

Several studies have shown substantial relations between monitoring ability and mathematical performance in primary school children (e.g., Desoete and Roeyers 2006; Desoete et al. 2001; Lucangeli and Cornoldi 1997; Özsoy 2011) as well as in secondary school children (e.g., Chen 2003; Roderer and Roebers 2013; Tobias and Everson 2000). However, in some instances, only low (Desoete 2008) or no associations between judgments and performance were found (e.g., Lucangeli and Cornoldi 1997). Since tasks differ regarding their demand on metacognitive monitoring and regulation processes, with highly routinized tasks requiring only little metacognitive regulation, this is not surprising. However, as will become apparent below, differences in the measurement of monitoring abilities that were used may also have caused the variability in findings.

1.2 Measurement of monitoring

A classic monitoring measure is calibration, which assesses the accuracy of metacognitive monitoring judgments by evaluating the fit between judgment and performance (Keren 1991). Thus, calibration combines two variables, namely, a judgment that predicts or postdicts performance in a cognitive task as well as the actual performance on this task.

Task performance is commonly measured categorically, being either correct or incorrect. Similarly, metacognitive judgments are often measured in a dichotomized form, judging task performance as correct or incorrect (Schraw et al. 2014). Aggregated across all items of a given test, judgment and performance data can be arranged in a 2×2 contingency table (see Table 1).

In Table 1, cell A contains the number of items that are judged as correct and solved correctly; Cell B contains items that are judged as correct, but solved incorrectly; Cell C contains items judged as incorrect, but solved correctly; Cell D contains items judged as incorrect and solved incorrectly. Consequently, cells A and D represent good calibration, whereas cells C and B indicate poor calibration. Cell C informs about the frequency of underconfidence, and cell B indicates the frequency of overconfidence.

As noted by Schraw et al. (2014), different statistical measures have been used to combine information contained in the four cells of the contingency table. Table 2 gives an overview of relevant constructs and statistical measures that are typically calculated based on the contingency table. While relative accuracy represents the ability of an individual to discriminate between items solved correctly and items solved incorrectly, absolute accuracy matches the judgments

Table 1 A 2×2 contingency table illustrating the performance-judgment array for monitoring accuracy (after Schraw et al. 2014)

	Performance		Row marginals
	Correct	Incorrect	
Judgment			
Correct	A	B	A + B
Incorrect	C	D	C + D
Column marginals	A + C	B + D	A + B + C + D

Table 2 Constructs and measures of metacognitive monitoring (adapted from Schraw 2009 and Schraw et al. 2014)

Construct	Measure	Formula	Description	Range	Interpretation
Absolute accuracy/calibration					
	Absolute Accuracy index (AAI)	(A + C) − (A + B)	Difference between actual correctly solved problems and problems judged as correct	− n–n	Perfect accuracy: 0
	Hamann coefficient (HAC)	((A + D) − (B + C))/(A + B + C + D)	Difference between the proportion of concordant and discordant judgments	− 1–1	Perfect accuracy: 1
Relative accuracy/resolution					
	Gamma (GMA)	(AD − BC)/(AD + BC)	Difference between product of concordant and discordant judgments	− 1–1	Perfect accuracy: 1
	d´ (DIS)	z (A/(A + C)) − z (B/(B + D))	Difference between standardized hit rate and false-alarm rate	− ∞–∞	Negative values: more false alarms than hits; positive values: more hits than false alarms
Diagnostic accuracy					
	Sensitivity (SEN)	A/(A + C)	Proportion of "I can solve" judgments when item is solved correctly (hit rate)	0–1	Perfect accuracy: 1
	Specifity (SPE)	D/(B + D)	Proportion of "I cannot solve" judgments when item is not solved correctly (correct-rejection rate)	0–1	Perfect accuracy: 1

n number of items judged

against the actual performance, representing an individual's ability to estimate performance on individual items.

In addition, Schraw et al. (2014) recommend indicators of diagnostic accuracy. Drawing on signal detection theory, they proposed sensitivity and specificity as measures of metacognitive monitoring. These measures discriminate between one's ability to judge items solved correctly (sensitivity) and items solved incorrectly (specificity). That is, they represent two complementary aspects of metacognitive monitoring, namely the identification of correct and of incorrect performance.

The variety of monitoring constructs evolved from the assumption that monitoring processes consist of different facets. Research testing this assumption is rare and inconclusive. Schraw et al. (2014) found substantial correlations between absolute, relative, and diagnostic accuracy measures. In particular, the correlation between absolute and relative accuracy measures was close to perfect (r's > .90). However, the correlation between sensitivity and specificity was close to zero. The authors concluded that measures of absolute and relative accuracy are indicators of the same monitoring processes, whereas sensitivity and specificity capture different processes. In contrast, Maki et al. (2005) reported low and nonsignificant correlations between absolute and relative accuracy measures (all r's < .15). According to Maki et al. (2005), this finding suggests that relative and absolute metacognitive accuracy measures tap into different processes.

Regarding prospective and retrospective judgments, most investigations focused on postdiction measures (e.g., Maki et al. 2005; Schraw et al. 2014), which, according to Bol and Hacker (2012), seem to be more accurate than predictions. In the field of mathematics, Boekaerts and Rozendaal (2010) confirmed this finding for arithmetic problems. However, retrospective monitoring accuracy decreased when students had to deal with word problems. Thus, current research on prospective and retrospective judgments is no more conclusive than research on the interrelations among absolute, relative and diagnostic accuracy measures.

1.3 Influence of response format on monitoring measures

1.3.1 Response format of the criterion

There is evidence that the format in which the criterion (i.e., the performance indicator) is answered affects the accuracy of metacognitive monitoring. For instance, Schwartz and Metcalfe (1994) reported higher accuracy scores for free recall (open-ended response format) than for recognition (closed response format). Their explanation is based on the impact of guessing. For example, suppose a student cannot solve a specific problem and knows it. In the case of an open-ended response format, his or her judgment should result in a correspondence between prediction and performance. In a closed format (e.g., multiple choice), however, the student most likely will guess, selecting one of the given alternatives. If he or she, by chance, guesses the correct solution, his or her actually appropriate judgment becomes incorrect.

Although such an explanation is tempting, it does not seem to apply to mathematics. Pajares and Miller (1997) presented the same word problems in an open- and in a closed-response format to middle-school students. Given the guessing option, multiple-choice items were easier than open-ended response items. However, contrary to Schwartz and Metcalfe's (1994) assumption, guessing did not corrupt the accuracy of the monitoring judgments. Students' judgments were actually more inaccurate in the open-ended response format than in the closed format. As students displayed a general overconfidence, lucky guesses increased calibration accuracy instead of decreasing it.

1.3.2 Response format of the judgment

Another factor that can influence monitoring accuracy is the scaling of the judgment, which determines the grain size of self-assessment. Typically, measurements of confidence include binary ratings (e.g., Tobias and Everson 2009), ordered categorical ratings (such as Likert scales; e.g., De Clercq et al. 2000) or, which is less common, continuous ratings (as visual analogue scales; e.g., Schraw et al. 1993).

An advantage of binary ratings (yes vs. no) is their direct correspondence to the binary performance scaling (correct vs. incorrect solution). As a result, calculation and interpretation of calibration measures are convenient. However, a disadvantage of binary ratings concerns the loss of information, as students may use more fine-grained internal categories for their confidence judgments than just yes or no (Higham et al. 2016).

Categorical and continuous scales show two advantages in comparison to binary scales: first, they can map nuances of confidence better; second, their distributional properties permit more sophisticated analysis options (e.g., structural equation modeling). Unfortunately, a crucial disadvantage of these judgment types concerns the divergence between confidence and performance scaling. There are three possible solutions to this problem: (a) dichotomizing the confidence rating scale and calculating a measure based on the contingency table, (b) modifying the performance rating scale and calculating a continuous calibration measure (Schraw 2009), or (c) calculating relative accuracy measures only. Whereas the first option discards information, the second option severely disregards the properties of the performance scale, and the third option constitutes a substantial restriction of analysis.

1.4 Influence of individual differences on monitoring measures

Regarding the influence of individual differences on monitoring indices, Bol and Hacker (2012) point out that "in general, higher-achieving students tend to be more accurate but more underconfident when compared to their lower-achieving counterparts" (p. 1). Two concurring processes may explain this phenomenon. On the one hand, abilities enabling a correct solution may be identical with those to predict or postdict the appropriateness of a solution. Thus, lower-achieving students may lack the knowledge required for appropriate metacognitive judgments as well as for adequate cognitive performance ("unskilled but unaware of it" as stated Kruger and Dunning 1999). Accordingly, higher-achieving individuals are more likely to judge their performance accurately. On the other hand, higher-achieving individuals tend to overestimate the mean solution rate, which leads them to underestimate their own capabilities (e.g., Dunning et al. 2003).

In the domain of mathematics, this effect has been only partially confirmed: comparisons between groups of differing achievement levels (García et al. 2016; Pajares and Miller 1997) as well as comparisons between problems of differing difficulty (Chen 2003) showed a general pervasive overconfidence. The degree of overconfidence, however, seemed to be a function of ability or task difficulty, with more accurate calibration for higher-achieving individuals or easier problems. Thus, in mathematics, accuracy seems to depend on performance in the criterion. However, in contrast to other domains there is little evidence for students displaying underconfidence.

1.5 Distributional effects on monitoring measures

Schwartz and Metcalfe (1994) pointed to the effect of restricted range in performance data. Given the fact that measures of calibration integrate judgments on correct as well as incorrect items, the range of difficulty in criterion tasks influences accuracy measures.

This phenomenon has been well illustrated in the special cases of ceiling or floor effects. For example, an overly difficult test leads to a reduced rate of correct criterion tasks (column marginal A + C in Table 1). Therefore—independently of monitoring competency—frequencies in cell A (correctly solved, judged as correct) and in cell C (correctly solved, judged as incorrect) are reduced. That is, all measures containing these two cells are biased. Due to the difficulty of the test, students do not have any chance to discriminate among these items, and the likelihood of observing a correspondence between judgments and performance is reduced. Therefore, comparing calibration scores between groups of different ability may confound monitoring proficiency and performance, even when the same test is used (Dunlosky et al. 2016; Schwartz and Metcalfe 1994).

Different calibration measures extract different information from the 2×2 table (see Table 2). Thereby, if cells in the table are empty, computational problems will emerge for some monitoring indices (Rutherford 2017). In short, empty cells can cause two problems, as follows. (1) Boundary values: quotients can result in a score of 1. Empirically, ceiling effects restrict variance. (2) Undefined values: the product of cell frequencies is 0, if one factor equals zero. In the case of the denominator this results in an undefined quotient. For example, gamma is not defined, if cells C or D are empty. Different features such as an insufficient number of items, extreme item difficulty, or extreme (in-) accuracy of judgments can lead to empty cells (Rutherford 2017).

1.6 Present study and research questions

It is obvious from the review of literature that methodological decisions influence monitoring accuracy scores. Due to the lack of systematic studies in the domain of mathematics, the impact of these methodological decisions on measurement results remains unclear. Our study aims at filling this gap by examining the effects of response and judgment format, ability level, and different calibration constructs in an ecologically valid mathematics education setting.

More precisely, we address seven questions: (1) Does response format influence performance? (2) Does type of judgment scale influence judgments and monitoring accuracy? (3) Does response format influence judgments? (4) Does calibration vary as a function of achievement level? (5) How does calibration inaccuracy affect the calculability and distribution characteristics of common calibration measures? (6) Do judgment scale and response format affect calibration measures in pre- and in postdiction? (7) Do calibration measures represent the same construct? Considering these questions may help researchers and practitioners to understand monitoring processes, to plan studies or evaluations, and to compare and integrate the literature in the field.

2 Methods

2.1 Participants and procedure

The sample consisted of 109 seventh-grade students (58% female students, mean age 148.7 months (SD = 4.7)), enrolled in five classes in the higher educational track (Gymnasium) of one school located in Germany. Research assistants administered instruments during two consecutive instruction periods (approximately 90 min) in the classroom. Within each class, research assistants assigned students randomly to one of four groups. Although the same mathematical tasks were administered in all groups, they differed with regard to solution formats (open-ended vs. multiple-choice, varied within each group) and judgment formats (Likert scale vs. visual analogue scale, varied between groups).

2.2 Instruments

Mathematics performance: The performance test consisted of 20 mathematical problems and was developed by the authors. The problems were based on the joint curriculum for secondary schools. The test contained 10 algebraic problems (terms and equations) and 10 word problems. Two examples are presented in the following. *Algebraic problem*: "$-2x = 3$, $x =$"; *Word problem*: "Marlene loves playing computer games. On Monday, she played for 2 h; on Tuesday, she played 1 h less. On Wednesday, she played twice as much as she played on Monday and Tuesday together. How many hours did she play overall?" Each student had to solve all 20 items, 10 of which were given in open- and 10 in closed-response format. Missings or indecipherable solutions were coded as incorrect. Cronbach's alpha of the performance test was .68, and the correlation with grades in mathematics amounted to $r = -.61$, indicating a sufficient criterion-based as well as curricular validity of the test.

Confidence ratings: Prediction was assessed by asking students to judge whether they would solve the problems correctly. Children were asked: "What do you think? Will you be able to solve the following problem?" Participants in two groups gave their ratings using either a 4-point Likert scale (LS) (no, surely not—likely not—likely yes—yes, most certainly) or using a visual analogue scale (VAS) with the poles "no, surely not" and "yes, most certainly". Ratings on VAS were measured in millimeters and transformed to a scale from 0 to 100. Cronbach's alphas for LS and VAS were .83 and .91, respectively. Postdictions were assessed immediately after having solved a problem. The questions were as follows: "What do you think? Did you solve the problem correctly?" As in the prediction situation, students delivered their ratings on a 4-point Likert scale (LS) (no, surely not—likely not—likely yes—yes, most certainly) or on a visual analogue scale (VAS) with the poles "no, surely not" and "yes, most certainly". Ratings on VAS were measured in millimeters and transformed to a scale from 0 to 100. Cronbach's alphas for LS and VAS were .88 and .82, respectively.

Grades in Mathematics: To obtain an achievement indicator that was independent of our metacognitive monitoring assessment, we asked students to provide for their grade in mathematics as stated in their last biennual report. In Germany, grades range from 1 to 6, with 1 indicating a very good achievement, and 6 indicating an insufficient achievement.

2.3 Design and analysis

The experimental design included two between-subject conditions (judgment scale and test configuration) and two within-subject conditions (item set and response format). Judgments of two subgroups were based on a Likert scale, and those of the other two groups on a visual analogue scale (judgment scale). Groups judging on the same scale either solved item set 1 in a closed-response format and item set 2 in an open-response format, or vice versa (test configuration; Table 3).

Accordingly, all students judged the same two sets of problems either as closed-response or as open-response problems (within-group variation). Thus, whereas between-subject comparisons could assess the effects of different judgments scales, within-subject comparisons could reveal the effects of response format in different item sets. Finally, the interaction of test configuration and item set could identify possible effects of response format in the same item set.

To test the assumed effects of the conditions simultaneously, we used two- and three-way ANOVAs with judgment scale, test-configuration, and/or item set as independent variables and performance, judgment, or calibration measure as dependent variables. Power analyses with GPower (Faul et al. 2007) revealed a power of .84 for between comparisons and .99 for within comparisons as well as the interaction of between and within factors for a sample size of 109 and medium effect sizes ($f=.25$; $\eta^2=.06$). For large effect sizes ($f=.40$; $\eta^2=.14$), power was .99 for all comparisons. The alpha-level used for these analyses was $p<.05$. Thus, for all comparisons of interest, there was more than adequate power to detect medium effect sizes.

2.4 Data preparation and analysis

To assess the calibration of judgments, we related prospective and retrospective judgments and performance. Since performance and judgments were assessed on different scale levels (continuous level for pre- and postdictions and categorical level for performance), we decided to dichotomize the judgments (see Sect. 1.3.2). We pooled two categories of the Likert scale ("no, surely not" and "likely not" means "no" resp. "0", ("likely yes" and "yes, most certainly" mean "yes" resp. "1"). Similarly, we bisected the visual analogue scale at 50 mm (judgments below mean "no" resp. "0", judgments above "yes" resp. "1"). To calculate calibration measures, we integrated both dichotomous measures in 2×2 contingency tables for pre- and postdictions. Omitted predictions (0.6%) and postdictions (6.8%) were coded as missing values. Items with missing pre- or postdiction judgments did not contribute to the 2×2 contingency table.

3 Results

3.1 Effects of response format on performance

In order to test the effects of response format on performance, a $2\times2\times2$ ANOVA (with the sum of correctly solved problems as dependent and response format, configuration, and judgment scale as independent variables) was conducted. Judgment scale ($F(1, 105)=0.74$, $p=.393$, $\eta^2=.01$) and configuration ($F(1, 105)=1.64$, $p=.203$, $\eta^2=.02$) did not show any impact on performance. The effect of item set was also not substantial ($F(1, 105)=5.34$, $p=.061$, $\eta^2=.03$). However, the interaction between item set and configuration was significant ($F(1, 105)=69.26$, $p<.001$, $\eta^2=.40$). Thus, the tests were equally difficult, no matter which judgment scale was rated. Given that test configurations were not significantly different, the allocation of items and response formats can be regarded as balanced. As expected, open-response items (set 2 in configuration A and set 1 in configuration B) were more difficult than closed-response items (see Fig. 1).

3.2 Effects of judgment scale on judgments

To examine the effects of judgment scale on prediction and postdiction judgments, judgments on closed- vs. open-response format (that is, the interaction between item set and configuration) were compared for Likert-scaled judgments and for visual-analogue-scaled judgments using two 2×2 ANOVAs.

For Likert scales, no differences in the prediction ratings for open and for closed items were found (interaction item set $\times$ configuration; $F(1, 55)=1.05$, $p=.310$, $\eta^2=.02$).

Table 3 Design of the study

		Test configuration	Performance response format		
			Closed response	Open response	n
Judgment scale	Likert scale	A	Item set 1	Item set 2	25
		B	Item set 2	Item set 1	28
	Visual analogue scale	A	Item set 1	Item set 2	29
		B	Item set 2	Item set 1	27

n sample size

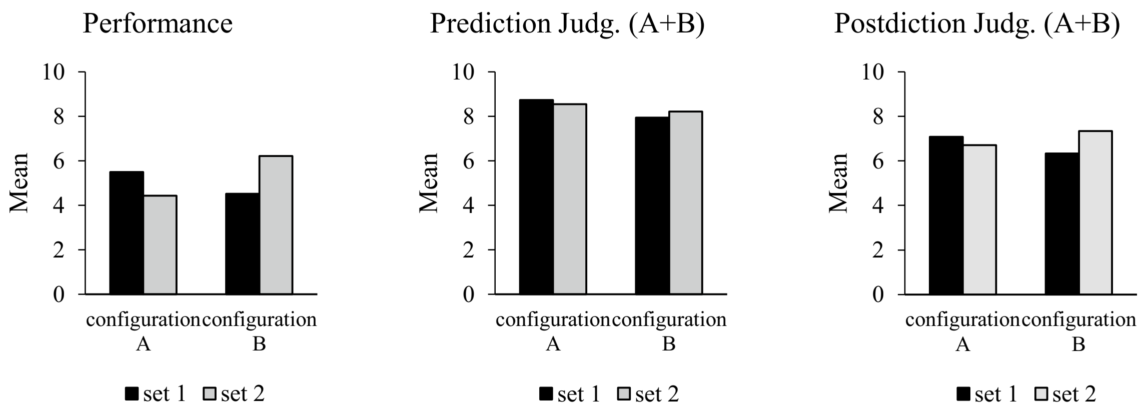

Fig. 1 Mean differences between groups and conditions

In contrast, students who rated their confidence on a visual analogue scale revealed different judgments for open and closed items, reporting less confidence for the open than for the closed items (interaction item set × configuration; $F(1, 50) = 7.78$, $p = .007$, $\eta^2 = .14$). Apparently, students used the fine-grained graduations of the visual analogue scale and, thus, discerned between response formats, whereas the 4-point Likert scale was not fine-grained enough to capture potential differences in confidence ratings.

In contrast to prediction, in postdiction students discriminated between response formats not only in visual analogue scaled judgments (interaction item set × configuration; $F(1, 50) = 36.99$, $p < .001$, $\eta^2 = .43$) but also in Likert-scaled judgments (interaction item set × configuration; $F(1, 55) = 36.58$, $p < .001$, $\eta^2 = .40$). For both types of judgments, confidence in open-response problems was lower than for closed-response problems. To compare pre- and postdictive confidence judgments, paired-samples t-tests were computed. In both judgment scale conditions, confidence judgments decreased from pre- to postdiction to a similar degree: Likert scale $t(55) = 7.56$, $p < .001$, $d = 1.04$; visual analogue scale $t(51) = 7.61$, $p < .001$, $d = 0.94$.

3.3 Calibration

To calculate commonly used calibration measures that draw on the 2 × 2 contingency table, we dichotomized pre- and postdiction judgments in the way described above. Tables 4 and 5 show mean frequencies of prediction and postdiction judgments, as well as performance scores. Regarding prediction (cf. Table 4), the majority of judgments is located in cell A (47%). Cell B contained 37% of predictions, indicating overconfident judgments. By contrast, judgments predicting failure were much more uncommon. Only in 11% of judgments failure was predicted correctly (Cell D). Cell C, indicating underconfidence, contained only 5% of the judgments and thus constituted the most infrequent category.

Table 4 2 × 2 contingency table for predictions

	Performance		
	Correct	Incorrect	Row marginals
Prediction judgment			
Correct	9.32 (47%)	7.39 (37%)	16.71 (84%)
Incorrect	0.99 (5%)	2.18 (11%)	3.17 (16%)
Column marginals	10.31 (52%)	9.57 (48%)	19.88

Table 5 2 × 2 contingency table for postdictions

	Performance		
	Correct	Incorrect	Row marginals
Postdiction judgment			
Correct	9.39 (50%)	4.28 (23%)	13.67 (73%)
Incorrect	0.87 (5%)	4.10 (22%)	4.94 (27%)
Column marginals	10.27 (55%)	8.38 (45%)	18.64

A comparison of Tables 4 and 5 reveals that the column marginals differ slightly. This is due to different rates of omitted pre- (0.6%) and postdictions (6.8%). Whereas the rank order of cell frequencies in postdiction remained constant, the relative frequency decreased in cell B (erroneously judged as correct, though wrong) and increased in cell D (accurately judged as wrong).

Paired-samples t-tests were used to compare pre- and postdiction contingency tables. To control for the differing total frequency of judged items, relative frequencies were compared. In cells A, B, and D significant changes occurred: $t(108) = 3.82$, $p < .001$, $d = 0.20$ (A); $t(108) = -10.13$, $p < .001$, $d = -0.94$ (B); $t(108) = 7.69$, $p < .001$, $d = 0.79$ (D). In Cell C, no change was found: $t(108) = -0.43$, $p = .671$, $d = -0.05$. The shift in cells B and D led to a closer correspondence between judgment and performance. However, students were still overconfident in their postdiction judgments (73% solution judged as correct vs. merely 55% of the solutions actually correct).

3.4 Effects of judgment scale and response format on calibration

As expected, the actual performance score was affected by response format, whereas the scale of prediction judgments did not seem to be relevant for performance. To test the pattern of effects for predicted performance, a 2 × 2 × 2 ANOVA (independent factors: judgment scale, configuration, item set) was calculated, with the row marginal A + B (sum of problems judged as solvable) as dependent variable.

Judgment scale ($F(1, 105) = 0.35$, $p = .558$, $\eta^2 = .00$), configuration ($F(1, 105) = 3.51$, $p = .064$, $\eta^2 = .03$) and item set ($F(1, 105) = 0.09$, $p = .760$, $\eta^2 = .00$) did not significantly affect students' "correct" judgments. The interaction between item set and configuration was also not significant, $F(1, 105) = 2.38$, $p = .126$, $\eta^2 = .02$. Thus, students did not align their prospective "correct" judgments on varying levels of difficulty. Rather, they judged open- as well as closed-response items overly optimistically, about to the same degree (see Fig. 1).

An analogous 2 × 2 × 2 ANOVA was computed for the retrospective "correct" judgments. Effects of judgment scale ($F(1, 105) = 0.00$, $p = .983$, $\eta^2 = .00$), configuration ($F(1, 105) = 0.35$, $p = .554$, $\eta^2 = .00$), and item set ($F(1, 105) = 2.30$, $p = .132$, $\eta^2 = .02$) remained nonsignificant. However, the interaction between item set and configuration reached significance ($F(1, 105) = 34.68$, $p < .001$, $\eta^2 = .25$). Thus, in contrast to prospective judgments, students were able to consider the varying levels of difficulty and judged their performance retrospectively in closed-response items more optimistically than in open-response items (see Fig. 1).

3.5 Calibration and achievement level

The 2 × 2 contingency tables indicate overconfidence for both predictions (Table 4) and postdictions (Table 5). Bias, that is, mean difference between "correct" judgments (cells A + B) and correct solutions (cells A + C), accounted for 6.4 items in prediction (SD = 4.4, range = − 7–16) and 3.4 items in postdiction (SD = 3.0, range = − 5–13), respectively. The decrease from pre- to postdiction was significant, ($t(108) = 7.33$, $p < .001$, $d = -0.72$), indicating a trend to more realistic judgments. Nonetheless, 84.4% of the students still overrated their performance on postdictions (predictions: 90.8%).

To explore the relation between bias and achievement level, we correlated the absolute values of bias scores and grades in mathematics. Grade and prediction bias were substantially and significantly correlated ($r = .426$, $p < .001$). The lower the achievement level of a student, the more pronounced was his or her prospective bias. For postdictions, there was no such relationship ($r = .115$, $p = .234$).

For a more detailed understanding of this finding, we inspected the degree of bias in different achievement groups. For this purpose, we split the sample by achievement level, using students' grades in mathematics as criterion variable. While there was no single student with a grade of 6, there were 5 students with a grade of 5, and 15 students with a grade of 4. We pooled these low achievers into a common group (grades 4 and 5). The resulting four achievement groups differed regarding actual performance scores ($F(3, 105) = 19.78$, $p < .001$, $\eta^2 = .36$) and mean postdiction scores ($F(1, 103) = 10.61$, $p < .001$, $\eta^2 = .23$). However, there was no difference between the groups regarding the mean prediction score, ($F(1, 103) = 1.37$, $p = .256$, $\eta^2 = .04$). Accordingly, there was a significant effect of achievement group on prediction bias ($F(1, 103) = 6.12$, $p < .001$, $\eta^2 = .15$), but not on postdiction bias ($F(1, 103) = 0.28$, $p = .840$, $\eta^2 = .01$).

As can be seen in Fig. 2, the predicted scores were approximately equal in all four achievement groups, though the actual scores differed across groups. Thus, prospective overconfidence increased with decreasing achievement level. The degree of retrospective overconfidence did not vary as a function of achievement level. Although the postdicted scores remained below the predicted scores, they still reflected overconfidence.

3.6 Calculability and distribution characteristics of calibration measures

Table 6 shows the descriptive statistics of calculated measures. Due to empty cells, gamma could not be computed for 26% (prediction) and 9% (postdiction) of the sample, with 46% (prediction) and 51% (postdiction) of scores showing a boundary value of − 1 or 1. In prediction, this is an effect of empty cells C and D. Students with missing gamma scores showed higher performance (column marginal A + C; $t(107) = 2.39$; $p = .019$; $d = 0.52$) and better grades in mathematics ($t(107) = -2.09$; $p = .039$; $d = -0.46$). For postdiction, this pattern was similar in that performance ($t(107) = 3.74$; $p < .001$; $d = 1.30$) and grades ($t(107) = -3.10$; $p < .001$; $d = -1.12$) was better for students with a missing gamma. Thus, missing values were not distributed at random. Gamma seemed to be systematically biased.

Measures of sensitivity indicated a high percentage of ceiling effects, with 53% (prediction) and 51% (postdiction) of the sample reaching the maximal value of 1. In the case of specificity, a bottom effect was found, especially for predictions, with 32% of the scores reaching the minimal value of 0. This tendency was also found—though to a lesser extent—for postdictions (10%).

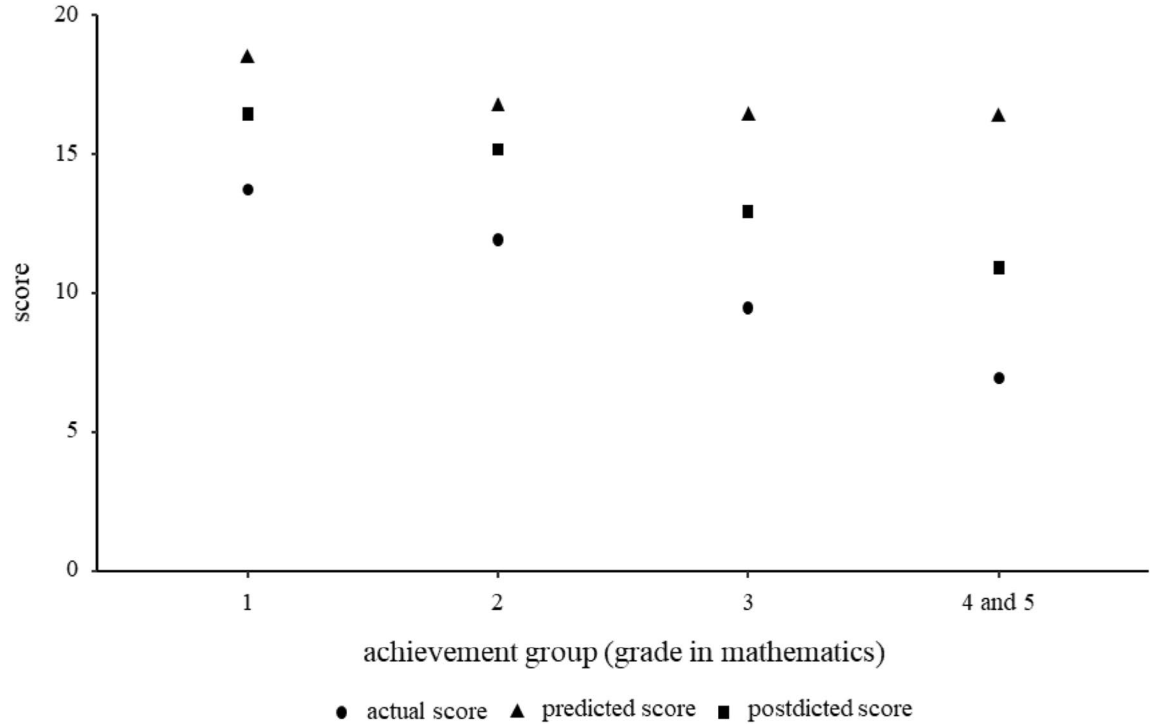

Fig. 2 Actual, predicted, and postdicted scores, as a function of achievement group

Table 6 Descriptive results for the various calibration measures

	N	Min	Max	M	SD	MD	MO
Prediction							
AAI	109	− 16	7	− 6.40	4.40	− 6.00	− 6.00
HAC	109	− 0.6	0.8	0.16	0.31	0.20	0.20
GMA	81	− 1	1	0.48	0.62	0.64	1.00
DIS	109	− 3.44	2.91	0.000	0.98	− 0.16	− 0.36
SEN	109	0.33	1	0.91	0.14	1.00	1.00
SPE	109	0	0.82	0.23	0.22	0.20	0.00
Postdiction							
AAI	109	− 13	5	− 3.40	3.04	− 4.00	− 4.00
HAC	109	− 0.3	0.9	0.45	0.24	0.50	0.50
GMA	99	− 1	1	0.82	0.29	0.97	1.00
DIS	109	− 2.75	2.13	0.00	0.95	0.06	− 1.06
SEN	109	0.29	1	0.90	0.14	1.00	1.00
SPE	109	0	1	0.45	0.26	0.50	0.50

N sample of calculable measures, *Min* minimal score, *Max* maximal score, *M* mean, *SD* standard deviation, *MD* median, *MO* modus

To investigate changes between pre- and postdiction, paired-samples t-tests were calculated for various measures of metacognitive monitoring (see Table 2). AAI (transformed in absolute values), HAC, GMA and SPE differed from pre- to posttest ($t(108) = 9.25$, $p < .001$ $d = 0.96$; $t(108) = -10.88$, $p < .001$, $d = 1.04$; $t(108) = -4.18$, $p < .001$, $d = 0.71$; t(108− 8.29, $p < .001$ $d = 0.94$, respectively), indicating an increasing accuracy from pre- to posttest. For SEN, no significant change could be found ($t(108) = 0.31$, $p = .760$). Given that DIS is standardized with $M = 0$ in pre- and in postdiction, a comparison was not possible.

3.7 Effects of judgment scale and response format on calibration measures

In order to test the effects of judgment scale and response format on calibration measures, we conducted 2 × 2 × 2 ANOVAs (judgment scale, configuration, response format). See Table 7 for a summary of results.

For predictions, we found two significant interactions between item set and configuration, indicating sensitivity for response format. The values of AAI were higher for the open- than for the closed-response version of the problems ($F(1, 105) = 22.80$, $p < .001$, $\eta^2 = .18$), indicating that open items led to a higher degree of bias. In line with this, the significant interaction between item set and configuration interaction observed for the HAC ($F(1, 105) = 12.28$, $p = .001$, $\eta^2 = .11$) consistently showed less accuracy for open items.

Additionally, analyses carried out for the AAI, HAC, and DIS indices indicated a small but significant effect of scale. The Likert scale led to more bias in the AAI ($F(1, 105) = 3.95$, $p = .049$, $\eta^2 = .04$), yielded a lower correspondence between judgment and performance in the HAC ($F(1, 105) = 6.97$, $p = .010$, $\eta^2 = .06$) and resulted in a smaller hit rate for the DIS ($F(1, 105) = 7.52$, $p = .007$, $\eta^2 = .07$).

For postdictions, a significant item set × configuration interaction was found for AAI ($F(1, 105) = 7.36$, $p = .008$, $\eta^2 = .07$) and HAC ($F(1, 105) = 10.09$, $p = .002$, $\eta^2 = .09$). The pattern of findings was roughly comparable to that of the predictions, though with reduced effect size. There were no significant differences between judgment scales.

3.8 Relations between calibration measures

The correlational pattern of measures shown in Table 8 indicates close relations among the calibration constructs assessing absolute accuracy (AAI and HAC, $r = -.82$ and $r = -.70$ for pre- and postdictions, respectively), and relative accuracy (GMA and DIS, $r = .78$ and $r = -.71$ for pre- and postdictions, respectively). In comparison, the correlation between sensitivity and specificity is lower but still substantial ($r = -.52$, Kendall's $\tau = -.42$, $p < .001$ for predictions and $r = -55$, Kendall's $\tau = -.43$, $p < .001$ for postdictions). The moderate but substantial correlations between measures of absolute, relative, and diagnostic accuracy point to common as well as unique psychological processes affecting monitoring.

The correlations between pre- and postdictions are either moderate (absolute accuracy), small (diagnostic accuracy), or nonsignificant (relative accuracy). This pattern of a rather low correspondence between pre- and postdictions even in the same measures points to different calibration processes.

4 Discussion

Regardless of the increasing number of studies that have shown the importance of metacognitive monitoring in mathematics education (see Baten et al. 2017 for a current review), there is only little empirical research on methodological issues in this domain. However, findings—predominantly from other domains—reveal that a student's monitoring skill is not only a function of ability but also of measurement choices. The experimental design of the present study permits researchers to investigate systematically the consequences of decisions that researchers and practitioners have to make before they measure students' monitoring skills.

First, we examined the accuracy of confidence judgments and their variation due to the resolution of judgment scale (Likert scale vs. visual analogue scale), the type of task response (open-ended vs. closed response), the phase of problem solving (pre- vs. postdiction), and the students' ability levels. Second, we analyzed calibration measures comparing indicators of absolute, relative, and diagnostic accuracy. We focused on issues that are especially relevant for research and practice in educational contexts, namely the susceptibility of the measures to boundary values (e.g., as a consequence of overconfidence), the impact of response format and judgment scaling on accuracy estimates, and the construct validity of different calibration measures.

4.1 Confidence judgments

Concerning confidence judgments, we first examined the impact of scaling. As expected, problems in the open-response condition were more difficult than in the closed-response condition. Students may have used the closed response format for lucky guesses or for a comparison of their own solution with the solution alternatives. Whereas judgments on the visual analogue scale mapped the differing difficulties in pre- as well as in postdictions, Likert-scaled judgments reflected them only in postdiction. Consequently, students are aware of differences in difficulty caused by response format, but a 4-point Likert scale seems to be too imprecise to map these differences, at least for performance prediction. Thus, researchers and practitioners interested in confidence judgments are advised to use visual analogues scales or at least finer grained Likert scales.

Second, we examined the overall accuracy of confidence judgments and effects of phase of problem solving. Although mean performance amounted to 52% correct solutions (ranging from 15 to 90%), judgments indicating

Table 7 Results of 2 × 2 × 2 ANOVAs

	Measure	Factor	Error df	df	F	p	η^2
Prediction	AAI	Item set	105	1	2.36	.128	.02
		Scale	**105**	**1**	**3.95**	**.049**	**.04**
		Configuration	105	1	2.96	.088	.03
		Item set × configuration	**105**	**1**	**22.80**	**<.001**	**.18**
	HAC	Item set	105	1	0.04	.844	.00
		Scale	**105**	**1**	**6.97**	**.010**	**.06**
		Configuration	105	1	1.90	.171	.03
		Item set × configuration	**105**	**1**	**12.28**	**.001**	**.11**
	GMA	**Item set**	**105**	**1**	**4.15**	**.046**	**.07**
		Scale	105	1	1.31	.257	.02
		Configuration	105	1	0.78	.382	.01
		Item set × configuration	105	1	2.73	.104	.05
	DIS	Item set	105	1	0.00	.953	.00
		Scale	**105**	**1**	**7.52**	**.007**	**.07**
		Configuration	105	1	0.12	.725	.01
		Item set × configuration	105	1	2.76	.100	.03
	SEN	Item set	105	1	0.59	.445	.01
		Scale	105	1	1.56	.215	.02
		Configuration	105	1	2.63	.108	.03
		Item set × configuration	105	1	1.47	.229	.01
	SPE	Item set	105	1	0.25	.616	.00
		Scale	105	1	2.51	.116	.02
		Configuration	105	1	2.18	.143	.02
		Item set × configuration	105	1	0.96	.328	.01
Postdiction	AAI	Item set	105	1	0.23	.635	.00
		Scale	105	1	0.56	.456	.01
		Configuration	105	1	0.96	.331	.01
		Item set × configuration	**105**	**1**	**7.36**	**.008**	**.07**
	HAC	Item set	105	1	0.47	.497	.00
		Scale	105	1	0.55	.460	.01
		Configuration	105	1	2.14	.147	.02
		Item set × configuration	**105**	**1**	**10.09**	**.002**	**.09**
	GMA	Item set	105	1	0.36	.552	.01
		Scale	105	1	0.89	.349	.01
		Configuration	105	1	0.00	.970	.00
		Item set × configuration	105	1	0.73	.397	.01
	DIS	Item set	105	1	0.00	.990	.00
		Scale	105	1	0.50	.483	.01
		Configuration	105	1	0.61	.473	.01
		Item set × configuration	105	1	0.70	.792	.00
	SEN	Item set	105	1	0.37	.547	.00
		Scale	105	1	0.05	.819	.00
		Configuration	105	1	0.27	.603	.00
		Item set × configuration	105	1	0.74	.391	.01
	SPE	Item set	105	1	1.35	.247	.01
		Scale	105	1	0.87	.354	.01
		Configuration	105	1	0.32	.574	.00
		Item set × configuration	105	1	0.29	.590	.00

df degree of freedom; significant effects are in bold

Table 8 Correlations among calibration measures

	AAI	HAC	GMA	DIS	SEN	SPE
AAI	**.46****	− .70**	.17	− .09	.38**	− 47.**
HAC	− .82**	**.53****	.40**	.47**	.17	.27**
GMA	.12	.25*	**− .03**	.71**	.53**	.14
DIS	− .14	.47**	.78**	**.14**	.48**	.48**
SEN	.42**	.12	.47**	.49**	**.28****	− .55**
SPE	− .56**	.34**	.40**	.49**	− .52**	**.30****

Above diagonal: postdiction scores; below diagonal: prediction scores. Diagonal: prediction with postdiction correlation (bold); * p < .05; ** p < .01

incorrect responses (cells C and D in the 2 × 2 contingency table) were scarce. For the majority of problems, students were confident in their ability to find or to have found the correct solution (cells A and B), confirming the pattern reported by Rutherford (2017). Moreover, students' overconfidence was stronger for prediction than for postdiction judgments, in which students did also in part align their overly optimistic assessment with the varying difficulty levels of open- and closed-response problems. The higher overall accuracy of retrospective judgments is in accord with the findings by Bol and Hacker (2012) and seems to be a consequence of intensive task experience (cf. Efklides 2008; Pressley and Ghatala 1990). Predictions had to be given after a brief exposure to the task, being based on a very short assessment of task requirements and a brief learning experience, and thereby possibly stimulating overly optimistic views regarding the outcome. Although judgments became more realistic and less overconfident from pre- to postdiction, it is important to note that overconfidence was dominating in both phases of the problem-solving process, confirming a robust phenomenon observed across various subject areas (e.g., Hacker et al. 2008; Nelson 1999). One implication of these findings is that seventh-graders experience problems distinguishing between difficult and easy tasks, or more exactly, identifying mathematics problems that they cannot solve. Therefore, educators should help students to acquire metacognitive knowledge regarding key task features and to implement monitoring and evaluation strategies using a variety of mathematics problems. To do so, educators should point out important task features as well as encourage students to check their understanding of a task before starting to work on it, and to evaluate the plausibility of local and final results.

Third, we examined the relationship between judgment accuracy and achievement level. In our sample, all achievement groups were overconfident. Predicted scores in the test were comparable across achievement groups, even though actual scores differed considerably. This pattern led to increasing overconfidence with decreasing achievement level. For postdictions, in contrast, overconfidence was at a comparable level for the four achievement groups and was generally much lower than for prediction judgments. Interestingly, lower-achieving students were able to evaluate their task performance quite accurately, thus showing about the same level of monitoring accuracy as higher-achieving students. These findings are only partly in accordance with the well-known impact of achievement level on confidence, indicating that higher-achieving students tend to show high accuracy but underconfidence, whereas lower-achieving students tend to show low accuracy but overconfidence (see Hacker et al. 2008). That is, at least for postdiction judgments, the "unskilled and unaware" effect (Kruger and Dunning 1999) may not be generalized to mathematics education. As our study indicates that even comparatively low-achieving students are able to evaluate their performance quite appropriately, we recommend encouraging students to evaluate their task solutions on a regular basis. This encouragement may help students to increase knowledge about task features as well as about their own strengths and weaknesses.

4.2 Accuracy measures

We compared indicators of absolute, relative, and diagnostic accuracy that are typically derived from the 2 × 2 contingency table (Schraw 2009). First, we examined the frequency of boundary values in an ecologically valid mathematics education context and their impact on calibration measures. Whereas measures of absolute accuracy were robust, gamma and the measures of diagnostic accuracy were sincerely biased. The main reason for boundary values (i.e., 0 or 1) and empty cells in the contingency table was the widespread overconfidence in our sample. As reported in the previous section, overconfidence seems to be a general judgment bias (e.g., see Rutherford 2017 for a similar result). Empty cells led to missing gamma values in up to one of four students, affecting particularly the higher achieving students. Thus, analyses in educational contexts with gamma may be biased. In the case of sensitivity and specificity, empty cells resulted for up to one of two students in boundary values, probably compromising the reliability of the measures. In practice, these findings indicate that gamma, sensitivity, and specificity should be interpreted with caution and may not be suitable for the assessment of metacognitive monitoring in natural educational contexts.

Second, we explored whether effects of response format and judgment scale on the cells of the 2 × 2 contingency table also affected calibration measures. We found that response format affected measures of absolute accuracy, with closed-response items eliciting more accurate monitoring than open-response format. Although this finding confirms the results of Pajares and Miller (1997) in the domain of mathematics, it is not in accord with the results reported by Schwartz and Metcalfe (1994) for metamemory. The decision to use an open- or closed-response format, which in research reports often gets lost in the shuffle of method sections, accounts for up to 17.8% of variance in absolute accuracy measures. The type of judgment scale affected prediction accuracy in absolute measures as well as in discrimination indices. Even after dichotomization, visual analogue scales led to a more accurate calibration, accounting for up to 6.7% of variance in prediction accuracy. In other words, ceteris paribus, response format as well as judgment scale impact calibration accuracy assessment substantially, especially in the case of absolute calibration measures. Thus, for analyzing and integrating research on monitoring accuracy, the measures used to assess monitoring need to be considered.

Third, we examined interrelations between measures of absolute, relative, and diagnostic accuracy to assess their convergent validity. As a main result, we found that correlations among measures reflecting the same type of accuracy were both statistically significant and substantial, whereas correlations among absolute, relative, and diagnostic accuracy scores were comparably low. Although this finding corresponds largely with the pattern reported by Schraw et al. (2014), our findings deviate from theirs to some extent. For instance, whereas Schraw and colleagues found that sensitivity and specificity tended to be uncorrelated, this was not true for our study where these two aspects of diagnostic accuracy were substantially interrelated. Thus, we could not confirm the assumption that sensitivity and specificity measure two independent calibration phenomena. However, our data confirm the assumption of Schraw et al. (2014) that indicators of relative and absolute accuracy may assess the same latent construct, albeit the correlations in our sample are much more moderate than Schraw's. A new aspect of our research concerns the rather modest correlations between the same accuracy measures assessed at different situations in the cognitive process (i.e., pre- vs. postdictions). With a few exceptions, these correlations were moderate to low, suggesting that different calibration processes took place in the two judgment conditions.

Given these findings on construct validity, we recommend that researchers should explicitly reference the selected accuracy construct. Furthermore, to improve the reliability of the measurement, we recommend computing and comparing at least two measures for each accuracy construct analyzed.

4.3 Limitations and directions for future research

Of course, the present study also suffers from some limitations. A first limitation is related to the sample selection. Only a small sample of secondary school students was recruited, all students were seventh graders, and all students attended the higher educational track of the German school system. Thus, it remains questionable whether our findings are representative for this age group and can be generalized to students attending lower educational tracks. Furthermore, given that no standardized mathematics test was available for seventh-grade students, a self-constructed test had to be used. A final limitation concerns the confinement to the mathematical domain, which impedes the drawing of conclusions regarding other domains such as monitoring of text comprehension.

In order to explore the generalizability of our findings, it seems important to replicate the study with a larger, more representative sample, and to extend the study goals. For instance, different age groups and tasks from different domains (e.g., mathematics and reading) should be included in the design to explore the robustness of findings. In addition, longitudinal designs and more elaborated statistical analyses such as latent growth modeling seem recommendable, in order to investigate whether calibration accuracy is an important predictor of further performance development. Finally, the impact of variables such as intelligence, self-concept, and motivation on calibration accuracy should be carefully assessed in future research. These variables may moderate calibration accuracy but may also be relevant when the goal is to change inappropriate metacognitive judgments.

References

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*(1), 3–38.

Baten, E., Praet, M., & Desoete, A. (2017). The relevance and efficacy of metacognition for instructional design in the domain of mathematics. *ZDM Mathematics Education, 49*(4), 613–623.

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.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology, 3,* 1–6.

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14,* 77–90.

De Clercq, A., Desoete, A., & Roeyers, H. (2000). EPA2000: A multilingual, programmable computer assessment of off-line

metacognition in children with mathematical-learning disabilities. *Behavior Research Methods, Instruments, & Computers, 32,* 304–311.
Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning, 3,* 189–206.
Desoete, A., & Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. *Learning and Instruction, 16,* 12–25.
Desoete, A., Roeyers, H., & Buysse, A. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities, 34,* 435–447.
Desoete, A., & Veenman, M. (2006). Metacognition in mathematics: Critical issues on nature, theory, assessement and treatment. In A. Desoete & M. Veenman (Eds.), *Metacogniton in mathematics education* (pp. 1–10). Haupauge: Nova Science.
Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–37). New York: Oxford University Press.
Dunlosky, J., & Tauber, S. K. (Eds.). (2016). *The Oxford handbook of metamemory*. New York: Oxford University Press.
Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87.
Efklides, A. (2008). Metacognition. Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13,* 277–287.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.
Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development* (4th ed.). Upper Saddle River: Prentice-Hall.
García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11,* 139–170.
Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16,* 163–176.
Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York: Psychology Press.
Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 39–61). New York: Oxford University Press.
Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134.
Lucangeli, D., & Cornoldi, C. (1997). Mathematics and metacognition: What is the nature of the relationship? *Mathematical Cognition, 3,* 121–139.
Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97,* 723–731.
Nelson, T. O. (1999). Cognition versus metacognition. *American Psychologist, 51,* 102–116.
Özsoy, G. (2011). An investigation of the relationship between metacognition and mathematics achievement. *Asia Pacific Education Review, 12,* 227–235.
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.
Polya, G. (1949). *Schule des Denkens—Vom Lösen mathematischer Probleme [How to solve it]*. Tübingen: Francke Verlag.
Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25,* 19–33.
Roderer, T., & Roebers, C. M. (2013). Children's performance estimation in mathematics and science tests over a school year: A pilot study. *Electronic Journal of Research in Educational Psychology, 11,* 5–24.
Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction, 47,* 33–42.
Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM—International Journal of Mathematics Education, 42,* 149–161.
Schoenfeld, A. H. (1985). *Mathematical problem solving*. New York: Academic Press.
Schoenfeld, A. H. (1987). What's all that fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 189–215). Hillsdale: Erlbaum.
Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415–429). Mahwah: Erlbaum.
Schraw, G., Kuch, F., Gutierrez, A. P., & Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology, 106*(4), 1192–1202.
Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*(4), 455–463.
Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge: MIT Press.
Tobias, S., & Everson, H. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147–222). Lincoln: Buros Institute of Mental Measurements.
Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 107–127). Mahwah: Erlbaum.