Acta Psychologica 220 (2021) 103399

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Exploring psychometric properties of children' metacognitive monitoring

Claudia M. Roebers [*], Mariëtte H. van Loon, Florian J. Buehler, Natalie S. Bayard, Martina Steiner, Eva A. Aeschlimann

University of Bern, Department of Psychology, Hochschulzentrum vonRoll, Fabrikstrasse 8, 3012 Bern, Switzerland

ARTICLE INFO

Keywords:
Metacognition
Monitoring
Children
Reliability
Stability over time

ABSTRACT

Two independent data sets assessing children's metacognitive monitoring abilities were used to explore the psychometric properties of classical and often-used monitoring measures in primary school age. Theoretically, monitoring is an overarching skill that helps individuals evaluate task mastery, strategy use, and correctness of performance. Monitoring skills are increasingly targeted when addressing individual differences in scholastic achievement and intervention approaches to foster students' self-regulated learning early on. In such contexts, knowledge about central psychometric properties is essential. Results of both studies revealed high internal consistency of prospective and retrospective monitoring judgments. When equivalent item sets (in terms of item difficulty) were considered (Study 1), split-half reliabilities were also satisfying. However, analyses revealed that the monitoring judgments' reliability depends on the reliability of the first-order task (recognition memory test). Retesting children of Study 2 after six months revealed considerable fluctuations in the monitoring measures. Among the included monitoring measures, reliabilities of within-person correlations (Gammas) between performance and confidence and recognition response times and confidence were poorest. Results are discussed in the context of the underlying theoretical construct and implications for research and practice.

For 40 years, researchers have investigated higher-order information processes in the context of learning and remembering, so-called metacognitive processes (Brown, 1978; Flavell, 1970). This traditional research is guided by the idea that learning and remembering rely on encoding, storing, and retrieving information and entails declarative knowledge and procedural skills. Procedural skills involve *how* to learn and remember and *how* to improve learning and remembering (Schneider and Löffler, 2016). Brown (1978) conceptualized procedural metacognition as self-regulatory and self-controlling information processes, including monitoring, planning, coping with failures, and evaluating performance. Thereby, monitoring skills are the starting point: by accurately judging how sure one is to remember information correctly, self-initiated and self-regulated learning and remembering becomes possible. Metacognitive monitoring processes thus require the ability to introspect and to self-evaluate performance, an ability now well documented to be emerging already in kindergarten children (Geurten et al., 2015; Ghetti et al., 2010; Lyons and Ghetti, 2013). The concept of metacognition is widely researched in cognitive, educational, and developmental sciences. However, very little information concerning the psychometric properties of classical measures of metacognitive monitoring is available. This is unfortunate because, increasingly, researchers investigate individual differences in children's monitoring and the effectiveness of monitoring interventions, and results from these studies are hard to interpret. Moreover, because of fatigue and low task persistence, the number of monitoring judgments in developmental studies is small compared to adult studies, impeding the possibility to generalize results from the general psychology literature. The present contribution aims to fill this gap in the literature, focusing on elementary school children's monitoring in a typical paired-associates learning paradigm with a subsequent recognition test.

Research has consistently shown that metacognitive processes are associated with learning, remembering, and academic performance in school-aged children. In a comprehensive review, Wang et al. (1993) concluded that metacognitive processes play one of the most critical roles in school learning, allowing the student to mastermind her or himself through any learning task. For example, reading comprehension (Artelt et al., 2001), writing (Hacker et al., 2009), mathematical problem solving (Lucangeli and Cornoldi, 1997; Rinne and Mazzocco, 2014), vocabulary learning (Pressley et al., 1987), and science learning (Van Loon et al., 2014) have been reported to be cross-sectionally and longitudinally linked to metacognitive monitoring skills. Typically, the effects of metacognition on memory and learning performance hold

[*] Corresponding author.
E-mail address: roebers@psy.unibe.ch (C.M. Roebers).

https://doi.org/10.1016/j.actpsy.2021.103399
Received 12 February 2021; Received in revised form 30 July 2021; Accepted 10 August 2021
Available online 25 August 2021

even after controlling for intelligence (van der Stel and Veenman, 2010; Veenman et al., 2005), underscoring the relevance of the metacognition construct. Given the strong impact of individual differences in monitoring for many aspects of scholastic achievement, it is all the more alarming how little is known about the psychometric properties of children's monitoring indicators.

Theoretically, metacognitive monitoring skills are considered an individual's "trait" (Efklides, 2011; Schneider, 2010; Wang et al., 1993). That is, one naturally assumes monitoring skills to be relatively stable over time. However, only a handful of studies has empirically investigated the intra-individual stability of metacognitive processes over time and across tasks. These studies have produced very inconsistent findings across a number of different indices of metacognition. In a recent study with second graders, 8-months-stability of children's correspondence of rated confidence and performance (bias score) was $r = 0.33$ (Roebers and Spiess, 2017). Rinne and Mazzocco (2014) reported that fifth to eighth graders' monitoring of mathematics performance was stable over a one-year delay ($r = 0.59$). It thus appears that the correspondence of performance and monitoring fluctuates over time and strongly depends on the study and the targeted age range. Children's ability to metacognitive differentiate between correct and incorrect performance, a measure of metacognitive discrimination or resolution, appears to fluctuate even more strongly over time. Two independent longitudinal studies (Roebers and Spiess, 2017; Steiner et al., 2020) focused on monitoring spelling and text comprehension, respectively. For both the participating second and fourth graders, the authors stated stability of metacognitive discrimination below $r = 0.20$.

One reason for the divergent findings may be the use of very different measures of metacognition. The bias score quantifies the absolute correspondence between performance and reported confidence (e.g., Lipko et al., 2009). In contrast, metacognitive resolution (intra-individual correlations between performance and confidence rating or difference scores) indicate to what extent an individual can accurately estimate the correctness of response (i.e., give higher ratings for correct and lower ratings for incorrect responses). This ability allows for the detection and correction of errors and adjustments of learning strategies (Flavell, 2000; Hembacher and Ghetti, 2014; Roebers, 2017). Another possible reason may lie in the fact that measures of children's monitoring stem from very different methodological approaches. Pronounced differences concern the first-order task and the test format (e.g., spelling task, math problems, learning and recalling picture pairs; open-ended questions vs. recognition; for example, Steiner et al., 2020) to which metacognitive processes are applied. Consequently, no matter how reliable the monitoring measures are, the differences in the first-order tasks are expected to lead to inconsistent findings across studies. Reporting psychometric properties has therefore never been of central interest.

Nonetheless, testing and reporting the most central and relevant psychometric properties of metacognition measures, especially internal consistencies and parallel test reliabilities, can help to estimate to what extent the different monitoring measures are affected by measurement error. For long, age-related group differences in metacognitive skills were investigated (Roebers, 2017; Schneider and Löffler, 2016; Schneider and Pressley, 1989). Moderate internal consistencies (around 0.40–0.50) are typically considered sufficient to document such group differences reliably (Lienert and Raatz, 1998) and thus did often not raise concerns. However, contemporary research increasingly addresses individual differences in metacognitive skills and how these differences relate to other variables of interest as well as intervention approaches to promote efficient self-regulated learning (e.g., school achievement; Dignath et al., 2008; Lyons and Ghetti, 2011; Pozuelos et al., 2019; Rinne and Mazzocco, 2014; Roebers et al., 2014; van der Stel and Veenman, 2010). When such questions are targeted, knowledge about the psychometric properties becomes viable for a clear-cut interpretation of the reported results. Unfortunately, prior studies have seldom described internal consistencies, parallel test reliabilities, or stability over time, especially in young samples.

1. The present study

We aimed to explore the psychometric properties of classical measures of children's metacognitive monitoring skills. In other words, we aimed to investigate how reliable and stable monitoring skills are. For this, we used a paired-associates learning task, in which second and fourth graders were asked to monitor the correctness of their recognition performance prospectively (i.e., before the memory test; judgments of learning, JOLs) and retrospectively (i.e., after the memory test; confidence judgments, CJs). We will report data from two studies. In Study 1, children completed two parallel versions of that learning task with a 10 min break in between. In Study 2, children within the same age range completed these two versions of the paired-associates task, but with a 6-months delay. For both data sets, we will report internal consistencies of the monitoring judgments (JOLs and CJs). While Study 1 additionally addresses the parallel test reliability of different monitoring measures, Study 2 will focus on the stability of individual differences in metacognitive monitoring measures over time.

We will present different measures of monitoring. For one, we will address the overall level of confidence concerning performance: children's bias or the realism of their monitoring by relating confidence to performance (Allwood, 2010; Baars et al., 2014; Howie and Roebers, 2007; Lipko et al., 2009; Rinne and Mazzocco, 2014). Positive bias scores indicate overconfidence, negative bias scores indicate underconfidence, and bias scores around zero represent realism. For another, we will investigate monitoring resolution, mirroring an individual's ability to metacognitively discriminate between correct and incorrect responses by contrasting monitoring judgments for correct with incorrect answers. Finally, drawing from the adult metacognition literature, we also include time-based monitoring measures (i.e., response times) as they are increasingly recognized as informative in children's studies. Experiences of more or less pronounced differences in response times (RT) during the memory test (i.e., retrieval fluency or memory vividness; short and longer response times during recognition) have been shown to be related to metacognitive monitoring in adults as variations of response times are used as memorial cues to inform monitoring (Koriat, 1997). Recent developmental findings uncovered that response times in a memory test are predictive for subsequent monitoring judgments cross-sectionally, and for improvements in memory performance longitudinally (Roebers et al., 2019). Against the background of such findings, it appeared fruitful to include within-person correlations between response times in the memory test and later confidence judgments (Flavell and Wellman, 1977). At the same time, we acknowledge the fact that these correlations are no pure measures of single cue use, as individuals have been found to use multiple cues, to use variable cues strategically, and that cues use and cue validity may be confounded (Bröder and Undorf, 2019; Undorf and Bröder, 2020).

Primary school children become quickly weary and typically show lower task persistence compared to adolescents and adults. Therefore, developmental researchers are forced to design child-appropriate tasks and, at the same time, psychometrically sound measurements. Shorter tasks with fewer items naturally come at costs for the psychometric properties but no study has yet systematically assessed psychometric properties of the classical monitoring measures in children. The primary aim was to present such data and analyze the relative quality of different monitoring measures derived from one task.

Based on the very few existing studies reporting psychometric properties of children's metacognitive monitoring, we expected relatively high internal consistencies. This is because researchers typically select homogenous items as first-order task material to which metacognitive processes will be applied (Lucangeli and Cornoldi, 1997). Research has also shown that monitoring judgments contain variance attributable to a "self-confidence" (personality) factor in children and adults (Dapp and Roebers, 2021; Kleitman and Gibson, 2011; Kleitman and Stankov, 2007; Roebers et al., 2012), leading to the expectation that individuals give similar judgments across items.

As to the parallel test reliability and stability over time, the evidence is so scattered that we did not have firm expectations. Monitoring resolution indicators (metacognitive discrimination and Gammas between performance and confidence) and time-based measures rely heavily on item characteristics (perceptual properties of a stimulus, semantic aspects, high vs. low associative item pairs; Dunlosky and Metcalfe, 2009). Therefore, we assumed that these measures have lower reliabilities and stabilities than trait-like measures like the bias score. Generally, and integrating Study 1 (10 min delay) and Study 2 (6 months delay), we anticipated that stability would be somewhat lower than the parallel test reliability. We anticipated this because the elapsed time between the two measurements (10 min vs. six months) should differentially influence the psychometric properties of individuals' developing monitoring skills in the targeted age groups.

2. Study 1

2.1. Method

2.1.1. Participants

The sample consisted of 206 children, including second graders (n = 93, 55% female; M = 100.8 months, SD = 5.5, age range: 89–113) and fourth graders (n = 113, 47% female; M = 129.4 months, SD = 5.1, age range: 121–144). Participants were recruited from public schools. Most participants were native speakers and of Caucasian ethnicity (71%); further 29% were non-native speakers (Eastern Europe's origin) who were sufficiently fluent in the local language to be regularly enrolled in school and participate in the study. A primary caregiver provided informed written consent; all children gave oral consent before testing. The ethics committee of the local faculty approved the study [approval number: 2016–0800004]. An additional 26 participants were tested but excluded from the analyses. Reasons for exclusion were technical problems (n = 6), inattention (n = 4), diagnosed attention deficit hyperactivity disorder or autism spectrum disorder (n = 5), lack of motivation for the paired-associates task (n = 2), no sufficient local language skills to understand the instructions (n = 2), age out of the targeted age ranges for second and fourth graders (n = 7).

2.1.2. Procedure and measures

Participants solved two parallel versions of a computer-based (Kanji) paired-associates task, Kanji-A and Kanji-B. Both versions were administered in one test session with a retest interval of 10 min in which participants solved riddles. They completed the Kanji tasks in groups of 10 to 25 participants at their schools. Altogether, the test session lasted approx. 65 min (instruction: 15 min; Kanji-A at T_1: 20 min; break/riddle-solving: 10 min; Kanji-B at T_2: 20 min). Two experimenters supervised participants and helped with the tablet computer if needed. Initially, the experimenters gave general instructions (what is a Kanji, how to use the touchscreen, and the Likert scale including example questions). Next, the experimenters introduced the task, and each participant completed a practice trial. During the task, the instructions were read aloud via headphones and appeared written on the screen. Participants gave answers by touching the screen twice (after touching once it was still possible to change the answer). Besides children's answers, the tablet computer also registered response times. The task consisted of different phases: study, delay, prospective global judgment, prospective monitoring (JOL), prospective control (restudy), recognition test and retrospective monitoring (CJ), retrospective global judgment, retrospective control (restudy and withdrawal of answers). In the present study focusing on monitoring, we used the JOL phase, recognition test, and CJ

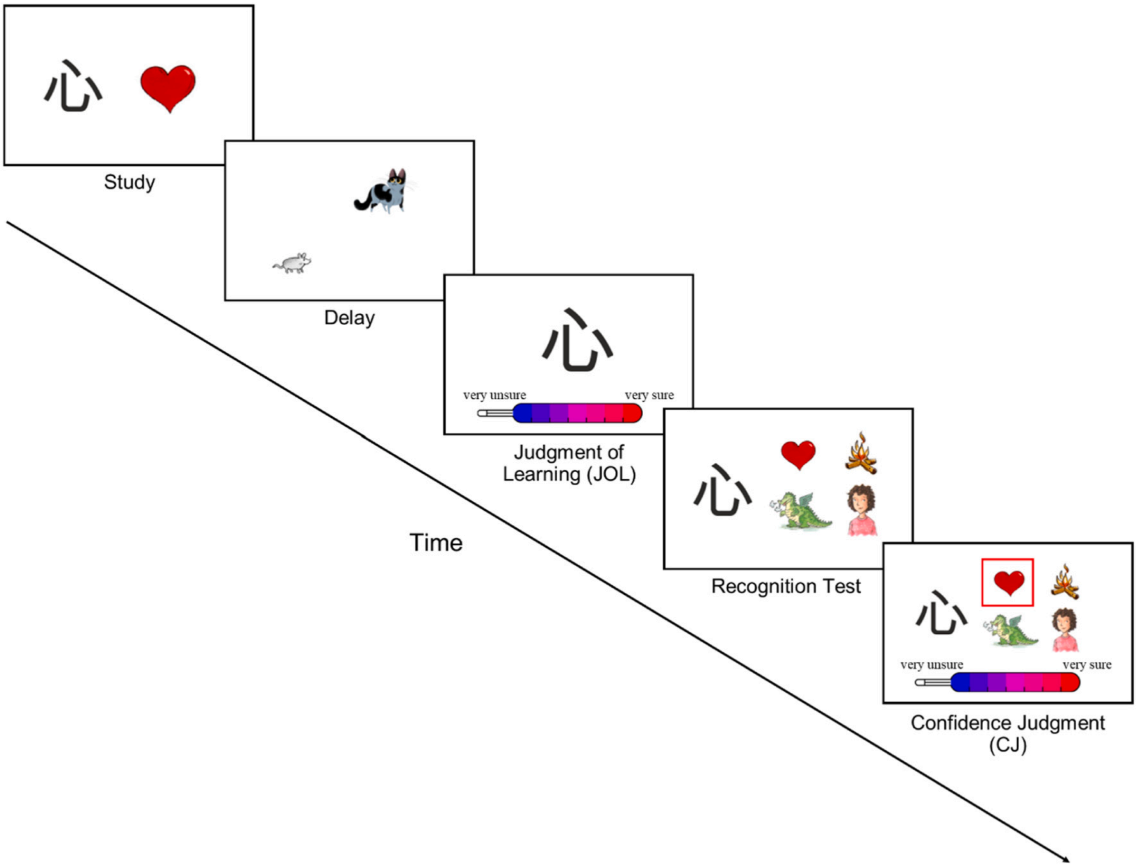

Fig. 1. Schematic illustration of the Kanji task procedure.
Note. In order to keep the illustration clear, instructions presented on the screen as text either bevor the start of the phase (which concerns study, delay and recognition test) or during the phase (which concerns JOL and CJ) are not depicted.

phase data. Hence, these phases plus the study phase are depicted in Fig. 1.

To begin with, participants learned item pairs, *Kanjis* (Japanese characters) and their corresponding meanings depicted as a color drawing (e.g., fire, mouth, or castle). Second graders learned 12 Kanjis, and fourth graders 16 Kanjis. Item pairs were presented each for 5 s in random order. Participants studied different item pairs in the Kanji-A version and the Kanji-B version to create two parallel task versions. Different item pairs were the only difference between the two versions. In advance, we had piloted a large pool of item pairs in order to obtain sufficient variability of item difficulties within one task and a comparable task difficulty index between both grades and task versions (difficulty index was between 0.11 (difficult) and 0.78 (easy)). Easier and more difficult items were not used for the present study as these have a too high likelihood to produce uniform monitoring data. After the study phase, participants solved a filler task on the computer (they had to catch a mouse with a cat; duration: 1 min) to prevent rehearsal or other memory strategies.

In the Judgment of Learning phase (JOL, prospective monitoring), each Kanji was presented, one at a time, together with a 7-point Likert scale, a colored thermometer (see Fig. 1). The scale ranged from 1 (very unsure) to 7 (very sure). Participants were asked the following question: "*How sure are you that you will find the correct picture for this Kanji?*". During the general instructions, the scale had already been introduced with example questions (e.g., *How sure are you that you know your teacher's name? How sure are you that you know my second forename?*).

In the Recognition Test and Confidence Judgment phase (CJ, retrospective monitoring), participants first had to select the correct meaning for the Kanji, and secondly, they were asked to give a CJ for this particular Kanji. Hereafter, the recognition test for the next Kanji appeared. In other words, each Kanji was presented together with four answer alternatives in the recognition test. All alternatives were drawings that had appeared in the study phase. Even if participants were unsure about the correct answer, they had to select one answer. Next, the selected answer was surrounded by a red frame, and the thermometer appeared at the bottom of the screen for the CJ. Participants were then asked the following question: "*How sure are you that you have chosen the correct picture?*". Thus, the CJ immediately followed the recognition of a single item.

2.1.3. *Dependent variables*

As dependent variables, we determined the accuracy of recognition in percent, different monitoring measures, and gamma correlations to quantify the associations between performance and confidence as well as the links between response times in the recognition test and confidence. These variables were calculated across all Kanjis (i.e., second graders: 12 Kanjis; fourth graders: 16 Kanjis). For the split-half reliability analyses, the values were calculated across only those 12 Kanjis solved by both age groups (see below which additional Kanjis were excluded for the split-half reliability). Otherwise, split-half reliabilities would not be comparable across the two age groups.

As to monitoring measures, we quantified children's bias in their monitoring judgments compared to performance, the confidence level for correct and incorrect, and the overall level of confidence (across all items, irrespective of recognition accuracy). Monitoring measures were calculated separately for the two monitoring judgment types, that is, JOLs and CJs. We computed the bias score by the following procedure: first, we defined the overall level of confidence in percentage for each participant. Second, we subtracted the recognition accuracy from the overall level of confidence. The bias score indicates either overconfidence (positive values), realistic evaluations (values around zero), or underconfidence (negative values), respectively (see, e.g., Baars et al., 2014; van Loon et al., 2013). Participants' ability to metacognitively discriminate (resolution) between correct and incorrect recognition was assessed by calculating the confidence level for JOLs and CJs as the average values based on either correctly or incorrectly recognized Kanjis, respectively (e.g., JOL $_{\text{correct}}$, JOL $_{\text{incorrect}}$).

Moreover, we used two different within-person gamma correlations for (a) quantifying the performance–confidence association (i.e., as an estimator for monitoring accuracy) and (b) the link between recognition response times and confidence judgments. The link between recognition response times and confidence judgments can hint towards children's cue use and has only very seldom yet been studied. Generally, Gamma correlations vary between −1 and +1, with higher values signaling a stronger association.

When calculating the Gamma correlations, some data naturally fell out of the analyses (uniform monitoring judgments; e.g., Wall et al., 2016). Before the gamma correlations with response times in the recognition test were computed, we removed extreme values of the response time variable by the following procedure: First, we deleted all responses $\leq$500 ms (as they are considered reflexes or corrections of the previous item). Second, we removed outliers if the response was outside ±3 *SD* from an individual's mean. Across all observations, we removed 0.4% of the data when computing descriptive statistics and 0.2% of the data when computing reliabilities. Note that gamma correlations cannot be computed for participants without any variability in confidence levels or recognition (e.g., 100% correct). Consequently, these participants dropped out of the corresponding computations. Of all within-person correlations, this concerned max. 5.0% of the data. Additionally, participants with very low variability in confidence judgments can get extreme gamma values (−1 or +1), falsely reflecting a perfect correlation. Hence, we also removed these values to obtain a more realistic estimation. Of all calculated Gamma correlations, this affected max. 11.2% of the data and is comparable to other studies (e.g., Wall et al., 2016).

2.1.4. *Statistical analysis*

We run the statistical analysis with SPSS (IBM SPSS Statistics, Version 25). For significance, we selected an α-level of 5% (two-tailed). As estimators of effect sizes, we will report partial eta^2 values ($\eta_p{}^2$) and correlation coefficients (*r*). In the descriptive statistics paragraph, differences between mean scores were examined by analysis of variances (ANOVAs) or multivariate ANOVAs (MANOVAs) with follow-up ANOVAs.[1] We tested whether the mean values of the bias scores and the Gamma correlations of the descriptive statistics data were significantly different from zero. For this, we calculated one-sample *t*-tests separately for both grades and both task versions. Because differences in mean scores are not in focus here, these results are presented in Appendix A.

To assess reliability in a first step, we assessed internal consistencies and parallel test reliabilities. In a second step, we computed split-half reliabilities. To quantify internal consistencies, we used the Kuder–Richardson Formula 20 (KR-20) for the dichotomous recognition variable and Cronbach's α for the confidence variables (JOLs and CJs) (Cronbach, 1951; Kuder and Richardson, 1937). The consistency coefficients can result in values from 0 to 1, with higher values representing higher internal consistency. To determine the parallel test reliabilities and the split-half reliabilities, we calculated partial correlations (*r*), controlling for age in months. For split-half reliabilities, we estimated the reliabilities for the whole test using the Spearman-Brown formula (Lienert and Raatz, 1998). To build the two test parts for calculating split-half reliability, we first excluded items with a very low (<20%) or very high (>80%) item difficulty index calculated with a formula correcting for chance (overall, four items were excluded). The difficulty index ranges from 0 to 100; the higher the value, the easier the Kanji item is. Next, we paired items with the same or a similar index and

[1] Depending of the analysis, we included two or more of the following independent variables: grades (second graders and fourth graders), task versions (Kanji-A and Kanji-B), monitoring type (JOL and CJ) and correctness of recognition (incorrect and correct).

randomly allocated them to one of the two test parts (Lienert and Raatz, 1998). To compute the equivalence between the two Kanji versions and between the two test parts, we applied the formula from Cureton (1971).

2.2. Results

2.2.1. Descriptive statistics

Table 1 presents the descriptive statistics of the percentages of correct recognition, the monitoring measures, and the gamma correlations as a function of age group (see Tables A1 and A2 in Appendix A for the results of the significant tests). Overall recognition accuracy averages 41%–66%, yielding an ideal database investigating monitoring data for correct and incorrect responses. As was expected, recognition accuracy was higher in fourth graders than second graders and higher in the Kanji-B than the Kanji-A task.

Concerning the bias score, Table 1 shows that fourth graders' monitoring was more realistic than second graders'. All mean values of the bias scores were significantly different from zero, except for fourth graders' JOL mean in the Kanji-A task (see Table A3 in Appendix A). These results show that second graders significantly overestimated their performance in both task versions; for fourth graders, this was only true for CJs. For JOLs, however, fourth graders gave a realistic evaluation in the Kanji-A task and even underestimated their performance in the Kanji-B task. Furthermore, independent of grade, children's monitoring was more realistic for JOLs than CJs, and more realistic in the Kanji-B than the Kanji-A task. In terms of confidence level, second graders' confidence was higher than fourth graders' confidence. Independent of grade, children were more confident retrospectively (CJs) than prospectively (JOLs), more confident in the Kanji-B than the Kanji-A task, and more confident for correct responses than incorrect responses. Gamma correlations between recognition and confidence levels of CJ (i.e., monitoring accuracy) were higher in fourth than in second graders. All gamma correlations were significantly different from zero (*ps* < .001).

Table 1
Descriptive statistics of recognition ACC, monitoring and gamma correlations of Study 1.

Variable	Task version	Second graders			Fourth graders		
		n	M	SD	n	M	SD
Recognition ACC (%)							
	A	93	41.04	16.22	113	59.57	18.45
	B	93	51.70	20.36	113	66.43	20.29
Bias score (% deviation) JOL							
	A	93	17.31	23.51	113	−2.35	20.06
	B	93	11.55	29.98	113	−6.80	22.84
CJ							
	A	93	26.15	20.17	113	8.23	20.10
	B	93	23.27	25.80	113	7.96	21.03
Monitoring (confidence level 1–7) JOL $_{\text{correct}}$							
	A	92	4.50	1.43	113	4.43	1.04
	B	92	4.72	1.46	113	4.48	1.21
JOL $_{\text{incorrect}}$							
	A	93	3.80	1.40	109	3.31	1.23
	B	93	3.96	1.71	108	3.48	1.34
JOL mean							
	A	93	4.08	1.34	113	4.00	1.04
	B	93	4.43	1.44	113	4.17	1.18
CJ $_{\text{correct}}$							
	A	92	5.18	1.40	113	5.27	0.97
	B	92	5.60	1.31	113	5.59	0.97
CJ $_{\text{incorrect}}$							
	A	93	4.35	1.58	109	3.88	1.29
	B	93	4.66	1.49	108	4.17	1.43
CJ mean							
	A	93	4.70	1.36	113	4.75	1.04
	B	93	5.25	1.20	113	5.21	1.00
Gamma correlations Performance-confidence							
	A	72	0.29	0.45	93	0.50	0.29
	B	65	0.38	0.44	87	0.50	0.36
Recognition RT-confidence							
	A	89	−0.29	0.30	107	−0.46	0.26
	B	83	−0.37	0.28	106	−0.40	0.29

Note. ACC = accuracy, JOL = judgments of learning, CJ = confidence judgments, $_{\text{correct}}$ = if recognition is correct, $_{\text{incorrect}}$ = if recognition is incorrect.

2.2.2. Reliability

As elaborated earlier, only a few existing studies report psychometric properties of metacognitive tasks. With Study 1, we aimed to investigate the reliability of the paired-associates task using two approaches: internal consistency and parallel test reliability. Because the parallel test reliabilities were relatively low, we additionally computed the split-half reliabilities.

The internal consistency signifies to what extent the different item-pairs of the task measure the same construct. Thus, for the two metacognitive monitoring types (JOLs and CJs), internal consistency constitutes an estimation to what extent a child tends to select low or high confidence levels consistently. Table 2 shows the KR-20 coefficient for recognition and Cronbach's α values for the confidence levels of JOLs and CJs. The internal consistencies were calculated for second graders and fourth graders, respectively, for the entire sample, for the two task versions, and both versions together. The values for recognition ranged from $r = 0.29$ to 0.73, indicating low to acceptable internal consistency. For the confidence levels, values were similar for JOLs and CJs and ranged from $\alpha = 0.81$ to 0.92. Hence, as expected confidence level's internal consistencies were high.

Table 3 presents the parallel test reliabilities and the split-half reliabilities. The parallel test reliabilities depict correlations between the two task versions. Reliability for recognition was low, with $r = 0.34$. Reliabilities for the monitoring measures were also low, with values ranging between $r = 0.41$ and 0.62. Bias scores and confidence level variables reached similar reliability values. Interestingly, reliability values of the gamma correlations were even lower and non-significant for the link between recognition response times and confidence.

The fact that internal consistencies of confidence were high but the parallel test reliabilities were low indicates that the two task versions were not entirely comparable. This was confirmed by the results of the

Table 2
Internal consistencies for recognition and confidence level of Study 1.

Variable	Task version	Second graders	Fourth graders	All Children
Recognition ACC (incorrect/correct)				
	A	0.29	0.62	0.60
	B	0.56	0.66	0.66
	A&B	0.55	0.73	0.73
Confidence level (1–7) JOL				
	A	0.89	0.81	0.85
	B	0.90	0.84	0.87
	A&B	0.92	0.89	0.91
CJ				
	A	0.89	0.82	0.86
	B	0.84	0.82	0.83
	A&B	0.91	0.88	0.90

Note. The reliability coefficients were calculated for recognition with the Kuder–Richardson Formula 20 and for the confidence level with Cronbach's α. JOL = judgments of learning, CJ = confidence judgments.

Table 3
Parallel test and split-half reliability of Study 1.

Variable	Parallel test reliability	Split-half reliability (with Spearman-Brown formula)
Recognition ACC (%)		
	0.34***	0.67***
Bias score (% deviation)		
JOL	0.41***	0.78***
CJ	0.54***	0.74***
Monitoring (confidence level 1–7)		
JOL $_{\text{correct}}$	0.50***	0.73***
JOL $_{\text{incorrect}}$	0.51***	0.84***
JOL mean	0.59***	0.90***
CJ $_{\text{correct}}$	0.43***	0.73***
CJ $_{\text{incorrect}}$	0.59***	0.77***
CJ mean	0.62***	0.87***
Gamma correlations		
Performance-confidence	0.33***	0.36***
Recognition RT-confidence	0.14	0.36***

Note. ACC = accuracy, JOL = judgments of learning, CJ = confidence judgments, $_{\text{correct}}$ = if recognition is correct, $_{\text{incorrect}}$ = if recognition is incorrect.

*** p < .001.

equivalence calculation, which revealed a relatively low equivalence of $r_{(\text{equivalence})} = 0.54$ for recognition, and 0.69 and 0.73 for the confidence level of JOL and CJ, respectively. One reason for the low equivalences may be that some children discovered strategies during the Kanji A task that they could apply in the Kanji B version (all children solved first the Kanji A and then the Kanji B version). This would also explain why children performed better in the Kanji B than the Kanji A version, although the item difficulty index, calculated in the pilot study, was the same for both task versions. Therefore, we reanalyzed the item difficulty index for the items used here. Results revealed a mean index of 48.5% for the Kanji B version and a mean index of 37.6% for the Kanji A version, verifying that the Kanji B version was easier than the Kanji A task. Therefore, we collapsed both task versions and calculated the split-half reliabilities. Compared to the equivalence values of the original Kanji A and B task versions, the equivalences between both test parts were excellent, with $r_{(\text{equivalence})} = 0.96$ for recognition; and 1 and 0.99 for the confidence levels of JOL and CJ, respectively.

The split-half reliability for recognition was adequate, with $r = 0.67$. Reliabilities for the monitoring measures were adequate to excellent, reaching values between $r = 0.73$ and 0.90 (see Table 3). Bias scores and confidence level variables reached similar reliability values. The highest reliabilities reached the JOL and CJ mean variables. Reliability values of the Gamma correlations were still very low, reaching values only between $r = 0.17$ and 0.36.

2.3. Discussion

Study 1 revealed satisfactory internal consistency of the different monitoring judgments that are likely attributable to individuals' tendency to give either lower or higher judgments, independent of age and independent of the correctness of the answer. The tendency not to use the full scale for the monitoring judgments might mirror a person's self-confidence or self-concept related to learning tasks in general (Dapp and Roebers, 2021; Kleitman and Stankov, 2007; Roebers et al., 2012). Motivation, task persistence, and general achievement level are additional candidate factors contributing to a relatively high uniformity of monitoring judgments. Since there were no substantial differences in internal consistency between the two age groups, the present study suggests that self-confidence (or self-concept) and other influential factors affect younger and older children's judgments in a similar way.

The low parallel test reliabilities are a worry because the number of items used for one task corresponds well to the number of items used in children's studies otherwise. The finding that creating two item sets comparable in difficulty leads to a marked increase in reliability (split-half) points to the crucial role of the numbers of items included and their difficulty. The results call for carefully tailoring experimental tasks for metacognition research in children. Researchers must find a good trade-off between measurement's reliability and children's fatigue and persistence.

Surprisingly, there were no systematic differences in reliability indicators between prospective and retrospective judgments. In the literature, prospective judgments-of-learning are often assumed to be of better quality as any memory test preceding monitoring judgments (as is the case for retrospective confidence judgments) will affect monitoring (Dunlosky and Metcalfe, 2009). Possibly children's monitoring is not yet well enough developed to yield a difference in reliability between JOLs and CJs, or the same factors similarly affect JOLs and CJs. Either way, the present study indicates that JOLs and CJs are equally reliable in young samples if item difficulties of the first-order task items are well balanced.

Study 2 had two principal aims. For one, we aimed to replicate findings from Study 1 in terms of the internal reliabilities of monitoring judgments. The identical item sets with an independent sample (of the same ages as in Study 1) were used. For another, we aimed to address the stability of monitoring skills over time. The only way to do this is to re-assess participants' monitoring after a considerable delay. Therefore, the identical two item sets of Study 1 were used in Study 2, however, with a delay of six months. As we were generally interested in whether age (and superior monitoring skills) affects psychometric properties of monitoring measures, we again included second and fourth graders.

3. Study 2

3.1. Method

3.1.1. Participants

The sample and data set of Study 2 were drawn from a larger longitudinal project for which developmental improvements in monitoring and control have already been reported (REFERENCE WITHHELD FOR BLINDED REVIEW). As in Study 1, the sample of Study 2 (n = 259) included second graders (n = 123, 52% female; M = 96.6 months at T_1, SD = 4.4, age range: 84–107), and fourth graders (n = 136, 49% female; M = 121.4 months at T_1, SD = 4.5, age range: 115–144). Participants were recruited from public schools, and most of them were of Caucasian ethnicity and native speakers (78%); 22% were non-native speakers (Eastern Europe's origin) with sufficient skills in the local language to participate in the study. Parents gave written informed consent; children gave oral consent before the testing session. Ethical approval from the study was obtained from the local faculty's ethics committee (approval number: 2016-08-00004). Sixty-eight participants were tested but excluded from the analyses. Reasons for exclusion were technical failures (n = 14), diagnosed attention deficit hyperactivity disorder or intellectual development disorder (n = 3), lack of motivation (n = 1), very low local language skills (n = 2), and chronological age falling out of our predefined range (n = 31). At $T_{2,}$ a new teacher refused participation, and thus seven children could not be tested (n = 7). A final ten children had moved out of the study's reach (n = 10).

3.1.2. Measures and procedure

The method and the two Kanji task versions were identical to Study 1, except that the interval between T_1 and T_2 was six months long. Hence, in Study 2, participants learned the same item pairs at T_1 (Kanji-A) and T_2 (Kanji-B). We determined the same dependent variables as in Study 1. Again, for the descriptive statistics, the variables were calculated based on all Kanjis (i.e., second graders: 12 Kanjis; fourth graders: 16 Kanjis); for the stability analysis, the variables were calculated across the 12 Kanjis solved by both grades to be able to compare the obtained values directly.

Before calculating Gamma correlations, we removed extreme values of response times of the recognition test as was done in Study 1 (we removed 0.4% of the descriptive statistics data and 0.2% of the stability data). Moreover, we had missing data in the gamma correlations due to no variability in monitoring judgments or recognition (of all correlations in Study 2, we had to remove max. 3.0%). In addition, we removed extreme Gamma values (−1 or +1), resulting from a very low variability in monitoring judgments (max. 8.6%).

3.2. *Results*

For Study 2, we ran the same statistical analysis as for Study 1, except for the above reported split-half reliabilities. We used the identical α-level (5%) and estimators of effect sizes (η_p^2 and *r*).

3.2.1. *Descriptive statistics*

Table 4 presents the descriptive statistics for both grades and at both measurement points of the percentages of correct recognition, the monitoring measures and the Gamma correlations (see Tables B1 and B2 in Appendix B for the results of the significant tests). Overall recognition accuracy averages between 45%–66%, providing an optimal database analyzing the metacognitive data separately for correct and incorrect responses. As expected, recognition accuracy was higher in fourth than in second graders and increased from T_1 to T_2.

Table 4
Descriptive statistics of recognition ACC, monitoring and gamma correlations of Study 2.

Variable	Time point	Second graders			Fourth graders		
		n	M	SD	n	M	SD
Recognition ACC (%)							
	T_1	123	45.12	17.56	136	55.70	15.97
	T_2	123	53.73	21.00	136	66.22	19.92
Bias score (% deviation)							
JOL							
	T_1	123	10.46	23.04	136	2.10	19.35
	T_2	123	5.37	25.16	136	−3.50	20.68
CJ							
	T_1	123	19.89	26.36	136	12.67	18.99
	T_2	123	14.15	23.10	136	6.57	19.78
Monitoring (confidence level 1–7)							
JOL $_{\text{correct}}$							
	T_1	123	4.29	1.29	136	4.41	1.12
	T_2	123	4.51	1.24	136	4.62	1.15
JOL $_{\text{incorrect}}$							
	T_1	123	3.56	1.28	135	3.60	1.17
	T_2	120	3.70	1.33	127	3.68	1.34
JOL mean							
	T_1	123	3.89	1.19	136	4.05	1.10
	T_2	123	4.14	1.14	136	4.39	1.15
CJ $_{\text{correct}}$							
	T_1	123	5.03	1.31	136	5.32	0.96
	T_2	123	5.25	1.22	136	5.57	1.09
CJ $_{\text{incorrect}}$							
	T_1	123	4.10	1.47	135	4.08	1.28
	T_2	120	4.05	1.44	127	3.87	1.56
CJ mean							
	T_1	123	4.55	1.31	136	4.79	1.03
	T_2	123	4.75	1.18	136	5.10	1.19
Gamma correlations							
Performance-confidence							
	T_1	101	0.36	0.35	124	0.48	0.33
	T_2	102	0.37	0.40	100	0.51	0.36
Recognition RT-confidence							
	T_1	118	−0.36	0.34	136	−0.44	0.26
	T_2	121	−0.36	0.31	127	−0.43	0.26

Note. ACC = accuracy, JOL = judgments of learning, CJ = confidence judgments, $_{\text{correct}}$ = if recognition is correct, $_{\text{incorrect}}$ = if recognition is incorrect.

Regarding monitoring, fourth graders' bias score was overall more accurate than second graders'. At both measurement points, second graders overestimated their performance, both prospectively and retrospectively. In contrast, fourth graders overestimated their performance only retrospectively (i.e., CJs at T_1 and T_2) and gave realistic evaluations prospectively (i.e., JOLs at T_1 and T_2). These bias score values were significantly different from zero, except for fourth graders' JOLs at both measurement points (see Table B3 in Appendix B for details of these results). Moreover, independent of grade, children's bias scores were more accurate prospectively (JOLs) than retrospectively (CJs), and the bias in children's monitoring decreased over time. Concerning the confidence level, independent of grade, children were more confident retrospectively (CJs) than prospectively (JOLs) and more confident for correct responses than incorrect responses. Gamma correlations of monitoring accuracy were higher in fourth graders than in second graders. All Gammas were significantly different from zero (*ps* < .001).

3.2.2. *Reliability*

In Study 2, we investigated reliability using internal consistency. Table 5 presents the KR-20 coefficients of recognition and Cronbach's α values of the confidence levels for JOLs and CJs. We calculated the internal consistencies for the second and fourth graders, respectively, and for the entire sample, both for the two measurement points separately and also across the two measurement points. For recognition, the KR-20 coefficients ranged from $r = 0.40$ to 0.69, indicating low to acceptable internal consistency, but the coefficients were higher at T_2 than T_1. For both monitoring types (JOLs and CJs), Cronbach's α was similar and ranged from 0.82 to 0.90, indicating high internal consistency for the tendency to report either lower or higher confidence levels.

3.2.3. *Stability*

Table 6 displays stability computed with partial correlations between the two measurement points after controlling for age in months. For recognition, stability over time was $r = 0.37$. For both monitoring measures, stability ranged from $r = 0.37$ to 0.54, with slightly higher values for confidence levels compared to the bias scores. Of the gamma variables, only the association between recognition response times and confidence was significantly stable over time with $r = 0.19$. As for the parallel test reliabilities reported in Study 1, an explanation for the low stability values here may be the rather low equivalence between the two Kanji task versions. In fact, the equivalence was $r_{(\text{equivalence})} = 0.68$ for

Table 5
Internal consistencies for recognition and confidence level of Study 2.

Variable	Time point	Second graders	Fourth graders	All Children
Recognition ACC (incorrect/correct)				
	T_1	0.40	0.40	0.44
	T_2	0.60	0.66	0.66
	T_1&T_2	0.63	0.67	0.69
Confidence level (1–7)				
JOL				
	T_1	0.87	0.86	0.86
	T_2	0.83	0.85	0.84
	T_1&T_2	0.90	0.89	0.90
CJ				
	T_1	0.88	0.82	0.86
	T_2	0.83	0.85	0.84
	T_1&T_2	0.90	0.88	0.89

Note. The reliability coefficients were calculated for recognition with the Kuder–Richardson Formula 20 and for the confidence level with Cronbach's α. JOL = judgments of learning, CJ = confidence judgments.

Table 6
Stability values between first and second measurement point of Study 2.

Variable	r
Recognition ACC (%)	0.37^{***}
Bias score (% deviation)	
JOL	0.37^{***}
CJ	0.41^{***}
Monitoring confidence level (1–7)	
JOL $_{\text{correct}}$	0.37^{***}
JOL $_{\text{incorrect}}$	0.47^{***}
JOL mean	0.54^{***}
CJ $_{\text{correct}}$	0.45^{***}
CJ $_{\text{incorrect}}$	0.45^{***}
CJ mean	0.53^{***}
Gamma correlations	
Monitoring accuracy	0.09
Recognition RT-confidence	0.19^{**}

Note. The correlations present partial correlations after controlling for age (in months). ACC = accuracy, $_{\text{correct}}$ = if recognition is correct, $_{\text{incorrect}}$ = if recognition is incorrect.
** $p < .01$.
*** $p < .001$.

recognition, and 0.63 and 0.62, for the confidence level means of JOL and CJ, respectively.

3.3. Discussion study 2 and general discussion

The current approach aimed to explore the psychometric properties of metacognitive monitoring measures in primary school children. While Study 1 addressed consistency of monitoring judgments, parallel-test, and split-half reliabilities, Study 2 was included to replicate internal consistencies of Study 1 and extend the scope to stability over time.

As to the internal consistency of monitoring judgments, both studies revealed sufficiently high consistency. This proved to be the case for prospective (judgments-of-learning, JoLs) and retrospective (confidence judgments; CJs). Moreover, internal consistency of monitoring was independent of age in the age range studied, as Cronbach's Alpha values did not vary systematically between second and fourth graders (see Tables 2 and 5). This result confirmed our expectations as a few previous studies have reported sufficiently high internal consistency of metacognitive monitoring judgments (e.g., Lucangeli and Cornoldi, 1997). Such findings are interpreted as showing that individuals have a certain tendency to give either higher or lower judgments, depending on their "self-confidence" or self-concept (Dapp and Roebers, 2021; Kleitman and Stankov, 2007). The present analyses provide convincing evidence that this self-confidence factor is already established in primary school children. This interpretation is supported by the fact that collapsing items across the two task versions slightly increased internal consistencies, independent of age (Study 1). Unfortunately, due to the long delay and the substantial developmental progression between the two measurements, it was not meaningful to collapse items from the two sets in Study 2.

The overall level of confidence, irrespective of the correctness of a given answer, is not necessarily diagnostic for a child's ability to metacognitively discriminate between correct (when sure or very sure confidence judgments would be accurate) and incorrect (when not sure or entirely unsure confidence judgments would be appropriate) responses. Yet, overall confidence is typically related to first-order task performance, is informative concerning a child's general self-evaluations, and has substantial motivational effects in the long-term. Therefore, the existence of a reliable measure for this monitoring aspect is of great value for research and practice alike.

In contrast to the good internal consistencies, Study 1 revealed low parallel test reliabilities. This result was expected as item specifics, such as the perceptual input and varying item difficulties (due to high and low associative item pairs) profoundly impact monitoring processes. Item specifics may be the most likely reason for the low parallel test reliabilities, especially when only a few items can be included in the analyses. Item specifics can also explain the age-independency of these results (Dunlosky and Metcalfe, 2009). On the one hand, monitoring recall of items with varying difficulties is important, as easy and difficult items provide anchors for being "sure" or "unsure". On the other hand, metacognitively differentiating between easy and difficult item pairs is not challenging, not even for primary school children, and thus also not diagnostic for individual differences in monitoring skills within homogenous age groups. Reliably capturing monitoring skills thus calls for a large enough number of items with medium difficulty. Although items of Study 1 and 2 had been drawn from a larger item pool after extensive piloting to construct two equally difficult task sets (A and B), follow-up analyses revealed that the two task versions had not been equivalent in these two particular samples. Collapsing across the two task versions and selecting items with equivalent (medium) item difficulties, and then calculating split-half reliabilities (consequently based on a larger number of monitoring judgments in Study 1), boosted reliability indices to an adequate level (Table 3). This series of analyses shed important light on the psychometric properties of metacognitive monitoring judgments: For one, they strongly depend on the psychometric properties of the first-order task to which metacognitive processes are applied (in our case memory recognition). For another, as item specifics are unavoidable and may vary from sample to sample, large enough item sets are necessary, allowing post-hoc exclusion of too easy and too difficult items that seriously impede the psychometric qualities of the most critical monitoring measures. Otherwise, addressing individual differences or assessing the effects of metacognitive training and interventions will be relying on unreliable estimators of children's monitoring skills (Dignath et al., 2008; Roebers et al., 2014), risking to produce non-replicable research findings.

Moreover, our detailed analyses uncovered that these split-half reliabilities were about equal when comparing prospective (JOLs) and retrospective (CJs) monitoring judgments. This outcome is surprising as children's retrospective judgments have repeatedly been reported to be more accurate (in relation to performance). Split-half reliabilities were also about equal when considering correct vs. incorrect first-order task performance monitoring. As accurate monitoring of incorrect responses typically poses more problems for children (Lyons and Ghetti, 2011; Roebers, 2017), the present results are of great importance. The measurement of uncertainty monitoring (monitoring incorrect performance) can be expected to be equally reliable as certainty monitoring, and young children's documented deficits therein are thus not reflections of unreliable measurements.

Study 2 included a longitudinal perspective allowing to address stability over time. While there is implicit theoretical consensus that metacognitive monitoring skills should be stable over time (Efklides, 2011; Flavell and Wellman, 1977), there is very little evidence for that claim. The present study confirmed what had been expected: stability over time was only moderate, independent of age, and independent of the monitoring variable. Considering that children's performance in the recognition test was even less stable over the six months delay, the moderate stabilities of the monitoring indices were satisfying. The not optimal equivalence of task version A and B probably compromised otherwise higher stability. It can thus be assumed that children with poor monitoring skills will – over time – slowly but surely fall behind in their ability to evaluate their own (academic) performance.

Predominantly in the adult and educational psychology literature, Gamma correlations associating single-item performance with monitoring judgments (monitoring accuracy) or linking monitoring judgments with choice latency in the first-order task (cue utilization) are used. We included these indicators based on within-participant correlations in the present studies. Both reliabilities (Study 1, Table 3) and stability over time (Study 2, Table 6) of these Gamma correlations were unacceptably low. These results confirm concerns raised against the use of Gamma correlations in research with children when only a somewhat limited number of items and monitoring judgments can be included (Roebers and Spiess, 2017). Children do not yet use the entire continuum of the confidence scale because they still have difficulties reporting on fine-tuned differences on the uncertainty end of the monitoring scale. However, such fine-tuned, varying judgments on a large item pool are necessary to obtain a more reliable estimation of monitoring skills with Gamma correlations. The present assessments of monitoring measures' psychometric properties can thus provide critical information for researchers.

Together, results from both studies uncovered psychometric strengths and weaknesses of the different monitoring measures in children. Because the task-bound nature of monitoring measures is unavoidable, researchers have to pay attention to the first-order task meticulously. With a large enough, carefully selected item pool, primary school children's monitoring judgments can be expected to be internally consistent and map individual differences in the ability to metacognitively discriminate between right and wrong. Future research should follow up on children's monitoring skills' stability over time as this issue is of great theoretical and practical relevance (Efklides, 2011; Roebers, 2017; Schneider and Löffler, 2016).

Acknowledgments

The Swiss National Science Foundation financed the present studies under Grant No. 100014_165477/1 awarded to the first author. We wish to thank the participating children, their parents and teachers, and the research assistants for helping with data collection. We also gratefully acknowledge the support of Jacob Raible and Stefan Kodzhabashev from the faculty's technology platform with programming the task.

Appendix A. Results of the significance tests of study 1's descriptive statistics

Table A1
Analyses of variance in recognition ACC and monitoring variables of Study 1.

DV	Effect	Direction	F(df)	η_p^2
Recognition ACC (%)				
	Grade	2nd < 4th	58.30 (1, 204)***	0.22
	Task	Kanji-A < Kanji-B	33.37 (1, 204)***	0.14
	Task × grade		1.57 (1, 204)	0.01
Bias Score (% deviation)				
	Grade	2nd < 4th	46.80 (1, 204)***	0.19
	Monitoring	JOL > CJ	167.42 (1, 204)***	0.45
	Task	Kanji-A < Kanji-B	4.63 (1, 204)*	0.02
	Monitoring × grade		1.81 (1, 204)	0.01
	Task × grade		0.40 (1, 204)	0.00
	Monitoring × task		8.39 (1, 204)**	0.04
	Monitoring × task × grade		0.28 (1, 204)	0.00
Monitoring confidence level (1–7)				
	Grade	2nd > 4th	4.99 (1, 194)*	0.03
	Monitoring	JOL < CJ	118.31 (1, 194)***	0.38
	Task	Kanji-A < Kanji-B	15.53 (1, 194)***	0.07
	Correct/incorrect	Correct > incorrect	368.84 (1, 194)***	0.66
	Monitoring × grade		0.54 (1, 194)	0.00
	Task × grade		0.45 (1, 194)	0.00
	Correct/incorrect × grade		13.49 (1, 194)***	0.07
	Monitoring × task		4.68 (1, 194)*	0.02
	Monitoring × correct/incorrect		11.57 (1, 194)***	0.06
	Task × correct/incorrect		0.14 (1, 194)	0.00
	Monitoring × task × grade		0.00 (1, 194)	0.00
	Monitoring × correct/incorrect × grade		2.61 (1, 194)	0.01
	Task × correct/incorrect × grade		0.93 (1, 194)	0.00
	Monitoring × task × correct/incorrect		1.20 (1, 194)	0.01
	Monitoring × task × correct/incorrect × grade		0.50 (1, 194)	0.00

Note. For the bias scores, the direction of the effects represents how accurate the judgment was. ACC = accuracy, DV = dependent variable, Monitoring = monitoring type (JOL vs. CJ), JOL = judgments of learning, CJ = confidence judgments, Correct/Incorrect = if recognition is correct/incorrect.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table A2
Multivariate analyses of variance (MANOVAs) and subsequent analyses of variance (ANOVAs) concerning the gamma correlations of Study 1.

Effect	MANOVAs			Subsequent ANOVAs			
	Pillai's V	*F*(3, 121)	η_p^2	DV (gamma)	Direction	*F*(1, 123)	η_p^2
Grade	0.08	3.56*	0.08				
				Performance-confidence	2nd < 4th	7.10**	0.05
				Recognition RT-confidence		3.65	0.03
Task	0.02	0.75	0.02				
Task × grade	0.04	1.71	0.04				

Note. The direction of the effects represents the strength of the correlation. Grade = second and fourth graders, Task = task version (Kanji-A vs. Kanji-B), DV = dependent variable.

* $p < .05$.

** $p < .01$.

Table A3
Results of one-sample *t*-tests testing whether the monitoring bias was different from zero in Study 1.

Monitoring type	Task version	Second graders				Fourth graders			
		t (92)	p	95% CI	r	*t* (112)	p	95% CI	r
JOL	A	7.10	<0.001	[12.5, 22.1]	0.59	−1.25	0.215	[−6.1, 1.4]	0.12
	B	3.71	<0.001	[5.4, 17.7]	0.36	−3.16	0.002	[−11.1, −2.5]	0.29
CJ	A	12.51	<0.001	[22.0, 30.3]	0.79	4.35	<0.001	[4.5, 12.0]	0.38
	B	8.70	<0.001	[18.0, 28.6]	0.67	4.03	<0.001	[4.0, 11.9]	0.36

Note. See Table 1 with the means and standard deviations of the monitoring bias scores. JOL = judgments of learning, CJ = confidence judgments, CI = confidence interval.

Appendix B. Results of the significance tests of the descriptive statistics of Study 2

Table B1
Analyses of variance in recognition ACC and monitoring variables of Study 2.

DV	Effect	Direction	F(df)	η_p^2
Recognition ACC (%)				
	Grade	2nd < 4th	35.39 (1, 257)***	0.12
	Time	$T_1 < T_2$	55.69 (1, 257)***	0.18
	Time × grade		0.56 (1, 257)	0.00
Bias score (% deviation)				
	Grade	2nd < 4th	12.90 (1, 257)***	0.05
	Monitoring	JOL > CJ	229.10 (1, 257)***	0.47
	Time	$T_1 < T_2$	16.62 (1, 257)***	0.06
	Monitoring × grade		0.90 (1, 257)	0.00
	Time × grade		0.02 (1, 257)	0.00
	Monitoring × time		0.31 (1, 257)	0.00
	Monitoring × time × grade		0.01 (1, 257)	0.00
Monitoring confidence level (1–7)				
	Grade		0.47 (1, 244)	0.96
	Monitoring	JOL < CJ	164.55 (1, 244)***	0.40
	Time		2.77 (1, 244)	0.01
	Correct/incorrect	Correct > incorrect	571.99 (1, 244)***	0.70
	Monitoring × grade		0.38 (1, 244)	0.00
	Time × grade		0.34 (1, 244)	0.00
	Correct/incorrect × grade		6.24 (1, 244)*	0.02
	Monitoring × time		2.40 (1, 244)	0.01
	Monitoring × correct/incorrect		58.55 (1, 244)***	0.19
	Time × correct/incorrect		8.85 (1, 244)**	0.04
	Monitoring × time × grade		0.01 (1, 244)	0.00
	Monitoring × correct/incorrect × grade		6.35 (1, 244)*	0.03
	Time × correct/incorrect × grade		0.24 (1, 244)	0.00
	Monitoring × time × correct/incorrect		6.25 (1, 244)*	0.02
	Monitoring × time × correct/incorrect × grade		1.14 (1, 244)	0.00

Note. For the Bias Score, the direction of the effects represent how accurate the judgment was. DV = dependent variable, ACC = accuracy, Monitoring = monitoring type (JOL vs. CJ), JOL = judgments of learning, CJ = confidence judgments, Correct/Incorrect = if recognition is correct/incorrect.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table B2
Multivariate analyses of variance (MANOVAs) and subsequent analyses of variance (ANOVAs) concerning the gamma correlations of Study 2.

Effect	MANOVAs			Subsequent ANOVAs			
	Pillai's V	*F*(3,174)	η_p^2	DV (gamma)	Direction	*F*(1, 176)	η_p^2
Grade	0.08	4.92**	0.08				
				Performance-confidence	2nd < 4th	11.33**	0.06
				Recognition RT-confidence		1.81	0.01
Time	0.01	0.73	0.01				
Time × grade	0.01	0.58	0.01				

Note. The direction of the effects represents the strength of the correlation. Grade = second and fourth graders, Time = measurement time point (T1 vs. T2), DV = dependent variable.

** $p < .01$.

Table B3
Results of one-sample *t*-tests testing whether the monitoring bias was different from zero in Study 2.

Monitoring type	Time point	Second graders				Fourth graders			
		t (122)	p	95% CI	r	*t* (135)	p	95% CI	r
JOL	T_1	5.04	<0.001	[6.3, 14.6]	0.41	1.27	0.208	[−1.2, 5.4]	0.11
	T_2	2.37	0.019	[0.9, 9.9]	0.21	−1.97	0.050	[−7.0, 0.0]	0.17
CJ	T_1	8.37	<0.001	[15.2, 24.6]	0.60	7.78	<0.001	[9.5, 15.9]	0.56
	T_2	6.79	<0.001	[10.0, 18.3]	0.52	3.87	<0.001	[3.2, 9.9]	0.32

Note. See Table 4 with the means and standard deviations of the monitoring bias scores. JOL = judgments of learning, CJ = confidence judgments, CI = confidence interval.

References

Allwood, C. M. (2010). The realism in children's metacognitive judgments of their episodic memory performance. In A. Efklides, & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 149–169). New York, NY: Springer.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363–383.

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*, 382–391.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica, 197*, 153–165.

Brown, A. L. (1978). *Knowing when, where, and how to remember: A problem of metacognition* (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cureton, E. E. (1971). The stability coefficient. *Educational Psychological Measurement, 31*, 45–55.

Dapp, L. C., & Roebers, C. M. (2021). Metacognition and self-concept: Elaborating on a construct relation in first-grade children. *PLoS ONE, 16*, Article e0250845.

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*, 101–129.

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

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

Flavell, J. H. (1970). Developmental studies of mediated memory. In *, Vol. 5. Advances in Child Development and Behavior* (pp. 181–211). Elsevier.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development, 24*, 15–23.

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail, & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Hillsdale, NJ: Erlbaum and Associates.

Geurten, M., Catale, C., & Meulemans, T. (2015). When children's knowledge of memory improves children's performance in memory. *Applied Cognitive Psychology, 29*, 244–252.

Ghetti, S., Castelli, P., & Lyons, K. E. (2010). Knowing about not remembering: Developmental dissociations in lack-of-memory monitoring. *Developmental Science, 13*, 611–621.

Hacker, D. J., Keener, M. C., & Kircher, J. C. (2009). Writing is applied metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 154–172). New York, NY: Routledge/Taylor & Francis.

Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*, 1768–1776.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*, 871–893.

Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences, 21*, 728–735.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161–173.

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

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und testanalyse [Testconstruction and testanalysis]*. Weinheim: Psychologie Verlags Union.

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*, 152–166.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*, 173–191.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154.

Lucangeli, D., & Cornoldi, C. (1997). Mathematics and metacognition: What is the nature of the relationship? *Mathematical Cognition, 3*, 121–139.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*, 1778–1787.

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development, 84*, 726–736.

Pozuelos, J. P., Combita, L. M., Abundis, A., Paz-Alonso, P. M., Conejero, Á., Guerra, S., & Rueda, M. R. (2019). Metacognitive scaffolding boosts cognitive and neural benefits following executive attention training in children. *Developmental Science, 22*, Article e12756.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*, 96–111.

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS One, 9*, 1–15.

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51.

Roebers, C. M., & Spiess, M. (2017). The development of metacognitive monitoring and control in second graders: A short-term longitudinal study. *Journal of Cognition and Development, 18*, 110–128.

Roebers, C. M., Cimeli, P., Rothlisberger, M., & Neuenschwander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: An explorative study on their interrelations and their role for school achievement. *Metacognition and Learning, 7*, 151–173.

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

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*, 2077–2089.

Schneider, W. (2010). Metacognition and memory development in childhood and adolescence. In H. S. Waters, & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 54–84). New York, NY: Guilford Press.

Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In J. Dunlosky, & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 491–518). New York, NY: Oxford University Press.

Schneider, W., & Pressley, M. (1989). *Memory development between 2 and 20*. New York: Springer Verlag New York Inc.

Steiner, M., van Loon, M. H., Bayard, N. S., & Roebers, C. M. (2020). Development of Children's monitoring and control when learning from texts: Effects of age and test format. *Metacognition and Learning, 15*, 3–27.

van der Stel, M., & Veenman, M. V. J. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences, 20*, 220–224.

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgments is strategic. *Quarterly Journal of Experimental Psychology, 73*(4), 629–642.

Veenman, M. V., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science, 33*, 193–211.

Wall, J. L., Thompson, C. A., Dunlosky, J., & Merriman, W. E. (2016). Children can accurately monitor and control their number-line estimation performance. *Developmental Psychology, 52*, 1493–1502.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294.