Journal of Applied Developmental Psychology 80 (2022) 101419

Contents lists available at ScienceDirect

Journal of Applied Developmental Psychology

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

The accuracy and annual rank-order stability of elementary school children's self-monitoring judgments

Mariëtte H. van Loon [*], Natalie S. Bayard, Martina Steiner, Claudia M. Roebers

Institute for Psychology, University of Bern, Switzerland

ARTICLE INFO

Keywords:
Self-monitoring
Global judgments
Development
Age differences
Rank-order stability

ABSTRACT

The present study investigated age-related development in children's metacognitive self-monitoring skills; eight-year-olds ($N = 140$) and ten-year-olds ($N = 164$) were compared. Children learned paired associates and completed a recognition test. Two types of monitoring judgments were compared: predictions and postdictions of performance. To investigate the rank-order stability of monitoring judgments, the task was repeated one year later. Prediction accuracy was low for both age groups and did not improve over time. Postdictions were more accurate than predictions; this indicates that self-test experiences support children to take actual performance into account when monitoring learning. For the second graders, postdiction accuracy improved over one year. Annual rank-order stability was found for predictions and postdictions, suggesting that habitual judgment tendencies affect children's monitoring judgments and judgment accuracy.

Theoretical models of self-regulated learning emphasize that self-monitoring while engaging in the learning process is a critical factor for studying effectively (Dunlosky & Rawson, 2012; Nelson & Narens, 1990; Rinne & Mazzocco, 2014; Schneider & Löffler, 2016; Thiede, Anderson, & Therriault, 2003). During the elementary school years, children are continuously confronted with self-monitoring challenges. For example, when completing school tasks, they need to estimate how well they are proceeding and decide when efforts are sufficient. Further, after practicing with self-tests and after completing school tests, they need to evaluate how well they performed to optimize their future learning (Schneider & Löffler, 2016). Accurate monitoring is beneficial for adults as well as children's learning (Dunlosky & Rawson, 2012; Rinne & Mazzocco, 2014). Without accurate self-monitoring, children might not spend enough time studying task materials they have not yet learned well (Schneider & Löffler, 2016). Considering the importance of self-monitoring for children's learning, it seems highly relevant to acquire insights into children's monitoring accuracy, and factors affecting children's self-monitoring. The present research aims to investigate this. Obtained findings may improve understanding of the development of children's self-monitoring skills, and bring practical insights for teachers aiming to support children with self-monitoring of learning.

As described by the metacognitive and affective model of self-regulated learning (MASRL model) by Efklides (2011), variable task factors and more stable person factors jointly seem to affect self-monitoring judgments (Dapp & Roebers, 2021; Efklides, 2011). At the task level, for instance, learning experiences (such as feelings of familiarity and perceived task difficulty) and retrieval experiences (such as experienced fluency when making a self-test) guide monitoring judgments (Dunlosky, Mueller, & Thiede, 2016; Koriat, 1997). At the person level, stable beliefs about competencies, such as self-esteem and self-concept appear to influence self-monitoring (Dapp & Roebers, 2021). The present research investigates how task and person factors simultaneously influence elementary school children's self-monitoring judgments.

Children's monitoring accuracy: Predictions and postdictions

Accurate self-monitoring seems a prerequisite for effective self-regulation of learning, and for achieving high learning performance (Rinne & Mazzocco, 2014; Schneider & Löffler, 2016). In previous studies, monitoring judgments were assessed both with predictions (i.e., judgments made after learning but before taking a test) and postdictions (i.e., judgments after taking a test). Notably, developmental processes affect children's predictive and postdictive monitoring judgments and their accuracy. For preschoolers and kindergartners (age 3–6), self-monitoring is rather inaccurate, such that relations between monitoring judgments and objective performance are weak, both for predictions (Lipko, Dunlosky, Lipowski, & Merriman, 2012; Lipko,

[*] Corresponding author at: Institute for Psychology, University of Bern, Hochschulzentrum vonRoll, Fabrikstr. 8, 3012 Bern, Switzerland.
E-mail address: mariette.vanloon@unibe.ch (M.H. van Loon).

https://doi.org/10.1016/j.appdev.2022.101419
Received 7 July 2020; Received in revised form 12 April 2022; Accepted 22 April 2022
Available online 5 May 2022

Dunlosky, & Merriman, 2009; Stipek, Roberts, & Sanborn, 1984) and postdictions (Van Loon & Roebers, 2017). After entering school, and especially between seven and ten years of age, children's ability to self-monitor their performance improves, yet remains far from perfect (Ghetti, Papini, & Angelini, 2006; Koriat, Ackerman, Lockl, & Schneider, 2009; Krebs & Roebers, 2012; Roebers, Mayer, Steiner, Bayard, & van Loon, 2019; Schneider & Löffler, 2016; Shin, Bjorklund, & Beck, 2007).

Age-related improvements in self-monitoring seem, at least partially, to be a result of natural, age-related development. When children mature, their memory, language, and problem-solving skills increase, and the maturation of the prefrontal cortex has been found to be associated with these improvements (Crone & Steinbeis, 2017; Kail, 2015). At the same time, metacognitive development seems, at least in part, due to schooling experiences (Dignath & Büttner, 2018; Lockl, 2010; Roebers et al., 2019). In school, cognitive and metacognitive experiences accumulate, and children's self-monitoring skills may benefit from their growing experiences (Bayard, van Loon, Steiner, & Roebers, 2021; Efklides, 2011). During the elementary school years, the demands on children's metacognitive skills gradually increase, as teachers expect children to become more independent when doing homework, seeking help, and preparing for tests (Collins, Brown, & Newman, 1988). Children thus increasingly gain experiences applying their monitoring skills to identify their current state of learning progress (Desoete, Roeyers, & De Clercq, 2003; Dignath, Büttner, & Langfeldt, 2008).

Improvement of monitoring accuracy over time may not be similar for predictions and postdictions. Hertzog, Saylor, Fleece, and Dixon (1994) found that for adults, particularly prediction accuracy may show intra-individual improvement, such that self-monitoring judgments become more accurate over time, after obtaining repeated experiences with a self-monitoring task. Hertzog et al. (1994) asked undergraduates to learn associated nouns and then predict their performance for a subsequent recognition test. After the test, participants made postdictions. One week later, the same participants completed the task again with different nouns. Correlations between predictions and performance (indicating monitoring accuracy) increased from 0.29 to 0.54 after one week (termed the "prediction upgrading effect", Hertzog et al., 1994), whereas correlations between postdictions and performance remained similar over time ($r = 0.59$ and 0.63, for the first and second measurement, respectively). This study thus suggests that adults' postdictions are generally more accurate than predictions and that predictions (but not postdictions) may show intra-individual improvement over time, as an effect of task and test familiarity.

However, some studies did not find that predictions improve over time, even when participants obtained repeated task experiences and insights into performance (Bol, Hacker, O'Shea, & Allen, 2005; Foster, Was, Dunlosky, & Isaacson, 2017; Hacker, Bol, & Keener, 2008). For instance, Foster et al. (2017) showed that college students' prediction accuracy did not improve across 13 exams over a semester, even when they were explicitly informed about performance. Also for children, repeated task experiences may not necessarily benefit prediction accuracy (Lipko et al., 2012; Shin et al., 2007). Lipko et al. (2012) asked six-year-olds and eight-year-olds to memorize 10 objects and then predict performance for a subsequent recall test. After receiving feedback about actual recall test performance, the procedure was repeated three times. Despite receiving feedback, prediction accuracy did not improve. Rather than being related to actual performance, predictions were mainly related to previous predictions.

In sum, despite developmental improvements in monitoring concerning the ability to discriminate single correct from incorrect responses, task-specific pre-, and postdictions are typically biased and judgment accuracy may be hard to improve (Lipko et al., 2009, 2012; Roebers, 2014; Schneider, Visé, Lockl, & Nelson, 2000; Stipek et al., 1984; Van Loon & Roebers, 2017). Particularly for children, monitoring judgments may not improve over time. Instead of reflecting actual insights into variances in learning and test performance, self-monitoring judgments may rather be based on a more general, stable judgment tendency. It is presently unknown to what extent children's predictions and postdictions are stable over time, rather than being affected by fluctuations in actual task performance. The present research aims to evaluate this.

Monitoring judgment stability

Research on the stability of personality characteristics suggested that individuals have a disposition to judge their performance in a constant way (Orth, Dapp, Erol, Krauss, & Luciano, 2020; Trzesniewski, Donnellan, & Robins, 2003). Specifically interesting for the present study is research addressing longitudinal judgment stability of self-esteem and self-concept. Self-esteem ratings indicate overall judgments of ability, whereas self-concept ratings reflect self-perceived performance in specific domains (Cole et al., 2001). Similar to monitoring judgments, self-esteem and self-concept judgments reflect knowledge and perceptions an individual has concerning her or his competencies and performance. As for self-monitoring measures, a trend toward overconfidence is typically reported for self-esteem and self-concept as well (Efklides & Tsiora, 2002; Stankov, Lee, Luo, & Hogan, 2012), and children who more accurately self-monitor performance also seem to have a smaller bias in self-concept (Dapp & Roebers, 2021).

Measures of rank-order stability give insights into the degree to which children's judgments maintain their relative ordering over time, regardless of group mean-level shifts in judgment magnitudes (Trzesniewski et al., 2003). For elementary school children, self-esteem and self-concept seem largely stable (Putnick, Hahn, Hendricks, & Bornstein, 2020; Trzesniewski et al., 2003). Specifically, for self-esteem, rank-order stability, as indicated with test-retest correlations, ranges between 0.39 and 0.46 (Trzesniewski et al., 2003); for self-concept, even higher rank-order stabilities have been reported, with test-retest correlations ranging between 0.84 and 0.95. This seems to indicate that these judgments are based on a rather general judgment tendency.

Measures of self-esteem and self-concept reflect self-perceived performance in a broad domain over a prolonged period, whereas self-monitoring judgments mirror one's insights into performance for a specific task at a particular moment in time. Further, monitoring judgments are directly comparable to task performance and therefore, measures of monitoring accuracy can be calculated, whereas such accuracy measures cannot be directly inferred for measures of self-esteem and self-concept. For monitoring judgments, research shows that adults, as well as school children's (age 9–12) judgments, are intercorrelated (Kleitman & Moscrop, 2010; Kleitman & Stankov, 2001, 2007). However, this research investigated correlations between judgments for different tests across task domains (e.g., general knowledge, math, perceptual tasks) at a specific moment in time; longitudinal data on rank-order stability of children's global monitoring judgments is still lacking.

A unique contribution of the present research is to provide insights into the longitudinal stability of children's task-specific monitoring judgments. By measuring monitoring judgments one year apart, we aimed to acquire insights into the annual rank-order stability of elementary school children's self-monitoring judgments. We presume that monitoring judgments may not only be related to one another at one moment in time but that, in line with findings from research on self-esteem and self-concept, children's monitoring judgments may also be related over a longer duration of time. If so, this could explain why monitoring judgments are, at least to some extent, resistant to task-specific variances in learning and task experiences.

When investigating effects of task and person factors on longitudinal judgment stability, it is particularly interesting to compare predictions with postdictions. Postdictions are typically more accurate than predictions, indicating that individuals are accounting for task and test experiences (Dunlosky & Metcalfe, 2009). If postdictions are more sensitive to variable task factors, it may be that these judgments are less based on more stable person factors (i.e., habitual judgment tendencies) than predictions, such that for postdictions, annual rank-order stability may be lower.

Present study

The present study investigated children's prediction and postdiction accuracy when self-monitoring their learning performance. Second and fourth graders completed a paired-associate learning task; they completed the same task with different task items again one year later. After learning associated pairs, children made predictions about their performance followed by a recognition test. Then, after completing the recognition test, children made postdictions.

Past research investigated monitoring accuracy either with item-by-item judgments which are made for each individually learned item or with global, task-specific judgments, mirroring an individual's evaluation of overall task performance (Dunlosky & Metcalfe, 2009). Most of the research on children's self-monitoring investigated the accuracy of item-specific judgments (see Schneider & Löffler, 2016, for an overview). However, also task-specific judgments are highly relevant for children, and these are common in everyday life. For instance, when children have multiple assignments to complete, they need to prioritize and plan how to allocate study time. To do so, they have to judge (i.e., make global predictions) how much they already accomplished and how much effort they still need to invest. Furthermore, after finishing schoolwork, children reflect on their learning and estimate how well they did (i.e., make global postdictions) to adapt future learning (Hacker, Bol, Horgan, & Rakow, 2000; Schunk & Zimmerman, 2012). Given the importance of task-specific global judgments for successful self-regulated learning, the present study particularly focuses on this judgment type.

Findings on the accuracy of children's task-specific monitoring judgments are still limited, and mainly based on studies that only investigated either predictions or postdictions, rather than comparing these. By comparing predictions and postdictions, we can obtain more detailed insights into how accurately children can self-monitor right after learning task materials, and whether providing them with a self-test may improve their monitoring accuracy. Moreover, to our knowledge, no research investigated the longitudinal rank-order stability of children's task-specific monitoring judgments. The present study aimed to fill this gap, by investigating the development of monitoring accuracy as an effect of age (second and fourth graders) and type of judgment (predictions and postdictions). By comparing two age groups and two measurement points, we aimed to address age differences cross-sectionally and longitudinally.

Monitoring accuracy was assessed with correlations between judgments and performance (cf. Connor, Dunlosky, & Hertzog, 1997; Dunlosky & Hertzog, 2000). In line with previous research (Koriat et al., 2009; Krebs & Roebers, 2012; Roebers, 2002, Roebers et al., 2019; Shin et al., 2007), we expected that both when considering predictions and postdictions, fourth-graders would be more accurate than second graders (Hypothesis 1). Although a direct comparison between prediction and postdiction accuracy of global monitoring judgments has not been reported for children, in line with research with adults (Connor et al., 1997; Dunlosky & Hertzog, 2000; Hertzog et al., 1994), we expected that postdictions would be more accurate than predictions for both age groups (Hypothesis 2). Moreover, we addressed whether intra-individual monitoring accuracy of predictions and postdictions improved over one year (Explorative Question 1).

It is yet unknown to what extent habitual judgment tendencies affect monitoring judgments over a longer delay, as monitoring stability has not yet been investigated longitudinally. The present research investigated judgment stability over one year; in line with research on general self-evaluations about competencies (Putnick et al., 2020; Trzesniewski et al., 2003), we expected significant rank-order stability for task-specific monitoring judgments (Hypothesis 3). This would bring evidence for longitudinally stable, trait-like judgment tendencies. Further, we exploratively addressed whether there are differences between predictions and postdictions in rank-order stability (Explorative Question 2).

The present research design enables us to address to what extent one's monitoring judgments are affected by habitual judgment and by actual task performance. Although research shows that predictions may be more strongly related to one's previous predictions than to actual task performance (Dunlosky & Hertzog, 2000; Hacker et al., 2000; Lipko et al., 2012), to our best knowledge, no research investigated the joint contribution of past judgments as well as actual performance. Therefore, we addressed this as Explorative Question 3. When judgments are particularly related to one's past judgments, this could imply that monitoring judgments are resistant to change. If judgments are mainly affected by actual task performance, this could imply that judgments are rather variable and potentially more easily improvable with classroom instructions and feedback.

Method

Participants and sampling

As we did not have clear indications of potential effects sizes, we could not conduct a-priori power analyses to determine the sample size. We anticipated documenting small to medium effect sizes and would need a sample of at least 120 participants per age group. To account for potential drop-out, a somewhat larger sample was recruited. The total sample at T_1 consisted of 304 children; 140 s graders (50% girls; $M_{age} = 7.6$ years; $SD = 0.5$) and 164 fourth graders (47% girls; $M_{age} = 9.6$ years; $SD = 0.5$). From T_1 to T_2, the sample attrition rate was 9.9% due to changes in residence and technical failures. Children were recruited from public schools with German as instructional language in the larger vicinity of a mid-sized university town in Switzerland. Sixty-six children (21.7%) were non-native speakers but had sufficient knowledge of the German language to attend regular classes and participate. None of the children spoke Japanese as a first or second language (they had no prior knowledge of the learning material). Note that even though at T_2 second graders transitioned to third grade and fourth graders to the fifth grade, participants will be referred to as second and fourth graders throughout, to facilitate presentation of the results.

Participating children were part of a larger study investigating monitoring and control for different types of tasks. In total, they participated in seven measurements over one year. For three measurements, Kanji materials were used, three measurements assessed text learning (for results, see, and during one measurement, classroom interactions between teachers and children were recorded (see Van Loon et al., 2021). For the present study, we compared the first Kanji measurement point with the last Kanji measurement point, which occurred one year later.

The research project was approved by the ethical review board of the Institute for Psychology, University of Bern. Parents had given written informed consent before the study. Before starting, children were told that they could withdraw without consequences at any time during the task. No child ever did.

Materials and procedure

Children were tested in school in small groups of six to 12 children. For the task, children learned Japanese characters (Kanjis) with their meanings. These Kanjis were used in previous studies and were found to be appropriate as learning materials for younger children, and suitable to detect age differences in monitoring accuracy (Destan, Hembacher, Ghetti, & Roebers, 2014; Destan & Roebers, 2015; Roderer & Roebers, 2010).

The materials and procedures were pilot tested with a different sample, to ensure sufficient variability in the difficulty of the items and to obtain comparable task difficulty for the two measurement points and the two age groups. Therefore, four school classes (a second grade, $N = 19$; third grade, $N = 21$; fourth grade, $N = 18$; and fifth grade, $N = 17$) were tested with a larger number of items (14 items for second and third graders; 18 items for fourth and fifth graders). For each measurement point and each age group, items with an item difficulty index (Moosbrugger & Kelava, 2008) between 0.11 (difficult) and 0.78 (easy) were selected. To obtain comparable levels of task difficulty, based on findings from the pilot study, the decision was made to have a different number of items for the second (12) and fourth graders (16).

At the start of each measurement point, children received general instructions from the experimenters and practiced the use of the monitoring scales. During the task, which was presented on a tablet computer, instructions were given orally via headphones and were additionally shown on the screen. Before starting the task, children completed a practice trial to become familiar with the task and the tablet computers. The task procedure is shown in Fig. 1.

Learning phase

Children learned 12 (second graders) or 16 (fourth graders) Kanjis and their meanings which were illustrated as colored drawings (see Fig. 1). The Kanji-picture associations were randomly presented for 5 s and were separated with a blank screen of 1 s. After learning the item pairs, a filler task was presented for 1 min (children played a game for which they tried to catch mice with a cat) to prevent rehearsal or other memory strategies.

Making predictions

After the learning phase, children predicted their overall recognition test performance. Second graders provided predictions on a 6-point scale and fourth graders on an 8-point scale, illustrated as a colored thermometer (adapted from Koriat & Shitzer-Reichert, 2002). The scales were subdivided into steps of two items so that the first segment of the thermometer corresponded to 0–2 items, the second segment to 3–4 items, the third segment to 5–6 items, and so on (see Fig. 1). To provide predictions, children responded to the following question: "*Later on you will have to select the corresponding picture for each Kanji. What do you think, for how many Kanjis will you find the right picture?*" To answer, children had to select a segment on the thermometer by touching the screen and had to confirm their response with a second screen touch. The presentation time of the scale was self-paced. At the start of the testing session, the scale was introduced with the use of different example questions.

Recognition test

After predicting their performance, children completed a recognition test. One Kanji at the time was presented on the left side of the screen together with four alternatives, including the correct answer (on the right side of the screen) depicted as colored drawings. All answer alternatives appeared in the learning phase and were randomly selected and presented in random order. The recognition test had a forced-response format. There was no time limit to complete the recognition test.

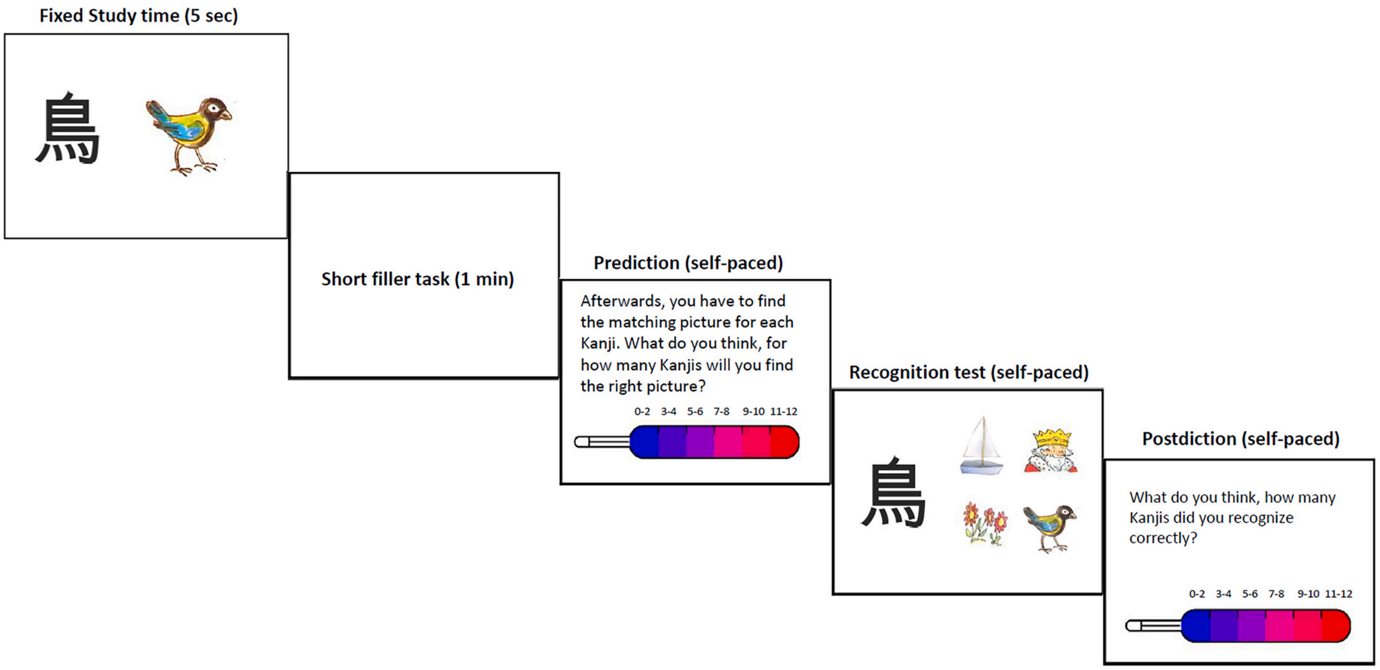
Making postdictions

Immediately after the recognition test, children had to make a postdiction by answering the following question: "*What do you think, how many Kanjis did you correctly recognize?*" by touching on the thermometer scale. Children estimated their performance with the use of the same scale as used for predictions. Again the presentation time was self-paced. At the end of the task, children were praised and were allowed to select a small gift.

Analyses

To assess children's recognition performance, the percentages of correct responses in the recognition test were calculated for second and fourth graders. To investigate the strength of the relation between performance and monitoring judgments, Pearson correlations were calculated between performance and predictions, and between performance and postdictions. Differences in correlations between age groups, differences between correlations for predictions and postdictions, and differences in correlations between measurement points, and were tested with the Fisher's r to z transformation.

To investigate the rank-order stability of monitoring judgments, partial correlations between monitoring judgments at T_1 and T_2 were calculated (controlling for recognition performance at T_1 and T_2), for the



Afterwards, you have to find the matching picture for each Kanji. What do you think, for how many Kanjis will you find the right picture?

What do you think, how many Kanjis did you recognize correctly?

Fig. 1. Task procedure.
Note. Example of the task procedure for second graders. Second graders learned 12 items, fourth-graders 16 items.

age groups separately. Differences between correlations for the age groups were again tested with Fisher's r to z transformation.

Regression analyses were used to investigate the effect of past monitoring judgments (made one year earlier at T$_1$), and actual recognition performance at T$_{2,}$ on monitoring judgments at T$_2$ (addressing Explorative Question 3). To make judgments and performance of second and fourth graders comparable, all predictors and outcome measures were transformed into z-scores per age group. Predictions and postdictions at T$_1$ as well as performance at T$_2$ were included as predictors of monitoring judgments at T$_2$. The interaction terms Age Group x Predictions at T$_{1,}$ Age Group x Postdictions at T$_{1,}$ and Age Group x Performance at T$_2$ were included to investigate potential age differences. Because literature is not clear about the relative importance of the predictors, all predictors and interaction terms were entered simultaneously.

Results

In the following section, we first report preliminary analyses of recognition performance, predictions, and postdictions. Then, we present analyses regarding our questions about monitoring accuracy and the stability of monitoring judgments for both age groups (second and fourth graders) and both judgment types (predictions and postdictions).

Preliminary analyses

Recognition performance

Table 1 shows descriptive statistics for recognition performance (both as absolute values and percentages) for both age groups and both measurement points. A mixed ANOVA with measurement point (T$_1$ vs. T$_2$) as within-subject factor and age group (second vs. fourth graders) as between-subjects factor revealed an improvement over time, $F(1, 272) = 123.52, p < .001, \eta_p^2 = 0.31$, and superior recognition performance for the older compared to the younger children, $F(1, 272) = 19.51, p < .001, \eta_p^2 = 0.07$. The Measurement Point x Age Group interaction was not significant, $F(1, 272) = 1.68$, $p = .196$, indicating that improvements over time did not differ between the two age groups.

Rank-order stability of recognition performance was measured with Pearson correlations between performance at T$_1$ and T$_2$. Correlations were low for second graders ($r = 0.19, p = .039$) and moderate for fourth graders ($r = 0.33$, $p < .001$), indicating that performance varied over time, particularly for the younger children. Values did not significantly differ between second and fourth graders ($z = -1.23, p = .110$).

Predictions

Table 1 shows descriptive statistics for both grades and both measurement points (in percentage). Second graders made predictions on a 6-point scale; both at T$_1$ and T$_2$, they expected to have 5–6 items correct out of 12 items. Fourth graders made predictions on an 8-point scale, they expected to have 7–8 items out of 16 items correct at T1 and 9–10 items correct at T$_2$. Overall, predictions were higher at T$_2$ compared to T$_1$ (second graders: $t(122) = 4.25$, $p < .001$, Cohen's $d = 0.39$; fourth graders: $t(150) = 5.03$, $p < .001$, Cohen's $d = 0.42$).

Postdictions

Table 1 shows the postdictions for both grades and both measurement points in percentage. Both at T$_1$ and T$_2$, second graders postdicted that they would have 5–6 items (out of 12 items) correct. For fourth-graders, the postdiction magnitudes indicate that both at T$_1$ and T$_2$, they expected to have 7–8 items correct (out of 16 items). Children in both age groups gave higher postdictions at T$_2$ than T$_1$ (second graders: $t(122) = -2.85$, $p < .001$, Cohen's $d = 0.25$; fourth-graders: $t(150) = -5.99$, $p < .001$, Cohen's $d = 0.52$).

Monitoring accuracy

Table 2 presents correlations between performance, predictions, and postdictions for both measurement points and both age groups. As can be seen in Table 2, the relation between predictions and performance was weak for second and fourth graders at both measurement points, indicating low monitoring accuracy, and no age differences were found in monitoring accuracy (T$_1$: $z = -1.56, p = .118$; T$_2$: $z = 1.57, p = .116$). Further, prediction accuracy did not increase from T$_1$ to T$_2$ for second graders ($z = -1.82, p = .069$) and for fourth graders ($z = 1.51, p = .131$). In summary, both age groups showed low accuracy when predicting their performance (in contrast to Hypothesis 1 expecting more accurate predictions for fourth than second graders), and there was no improvement in prediction accuracy over time (answering Explorative Question 1 for predictions).

Table 2 shows that for second graders, postdictions were not related to performance at T$_1$, whereas correlations between postdictions and performance were significant, indicating moderate monitoring accuracy at T$_2$. For fourth-graders, the correlations between postdictions and performance show moderate judgment accuracy at both measurement points. Hypothesis 1 was partially confirmed for postdictions; fourth graders showed better monitoring accuracy than second graders at T$_1$ ($z = -2.89, p = .004$), however, age groups did not differ in their accuracy at T$_2$ ($z = 0.87$, $p = .385$). Further, analyses answering Explorative Question 1 for postdictions show that monitoring accuracy improved for second graders from T$_1$ to T$_2$ ($z = -3.87$, $p < .001$), whereas no improvement in accuracy over time was found for fourth graders ($z = 0$, $p > .999$).

Hypothesis 2, suggesting that postdictions would be more accurate than predictions, was partially confirmed for second graders; there were no significant differences between these judgment types at T$_1$ ($z = -0.24, p = .811$), however, at T$_2$, postdiction accuracy was higher than prediction accuracy ($z = -3.26, p < .001$). Hypothesis 2 was confirmed for fourth graders; they were more accurate when making postdictions than predictions at both measurement points (T$_1$: $z = -2.36, p = .018$;

Table 2
Correlations between performance and monitoring judgments for both measurement points.

	1	2	3	4	5	6
1 Predictions T$_1$	–	0.57**	−0.01	0.41**	0.33**	0.10
2 Postdictions T$_1$	0.55**	–	−0.01	0.45**	0.49**	0.11
3 Performance T$_1$	0.16*	0.33**	–	−0.01	0.07	0.19*
4 Predictions T$_2$	0.43**	0.56**	0.16	–	0.63**	0.19*
5 Postdictions T$_2$	0.28**	0.53**	0.22**	0.66**	–	0.42**
6 Performance T$_2$	0.06	0.06	0.33**	0.00	0.31**	–

Note. * $p \le .05$; ** $p \le .01$. Values for second graders above the diagonal; values for fourth graders below the diagonal.

Table 1
Recognition performance, predictions, and postdictions for both age groups.

Grade	Recognition T$_1$ (%)*	Recognition T$_2$ (%)*	Predictions T$_1$ (%)**	Predictions T$_2$ (%)**	Postdictions T$_1$ (%)**	Postdictions T$_2$ (%)**
2	45.12 (17.56)	62.13 (20.67)	56.37 (21.18)	64.64 (18.30)	52.84 (22.55)	58.54 (21.16)
4	54.76 (15.92)	68.21 (19.71)	56.53 (20.10)	64.98 (18.32)	50.75 (18.83)	60.51 (21.94)

Note. Standard deviations of the mean in parentheses. *Second graders could recognize a maximum of 12 items; fourth-graders could recognize a maximum of 16 items. **Second graders made predictions and postdictions with use of a 6-point scale; fourth-graders made predictions and postdictions with an 8-point scale.

T_2: $z = -4.69$, $p < .001$).

Judgment rank-order stability

Partial correlations (controlling for performance at T_1 and T_2) between predictions at T_1 and T_2 showed significant judgment rank-order stability for second ($r = 0.40$, $p < .001$) and fourth graders ($r = 0.41$, $p < .001$); correlations did not differ between the two age groups ($z = -0.10$, $p = .918$). For postdictions, partial correlations between monitoring judgments at T_1 and T_2 (controlling for performance at T_1 and T_2) showed significant judgment rank-order stability for second ($r = 0.49$, $p < .001$) and for fourth graders ($r = 0.52$, $p < .001$), and there were no age group differences ($z = -0.33$, $p = .743$). This confirms Hypothesis 3 assuming significant rank-order stability between judgments, even when controlling for children's actual task performance. Analyses addressing Explorative Question 2 did not show differences between prediction and postdiction rank-order stability for both age groups (second graders: $z = -0.87$, $p = .192$; fourth graders: $z = -1.198$, $p = .115$).

Relations between judgments and performance over time

With regression analyses, we address Explorative Question 3 by investigating how monitoring judgments made at T_1 and recognition performance at T_2 affect monitoring judgments at T_2. As shown in Table 2, the predictors in the regression analyses were not highly correlated with each other, indicating that issues with multicollinearity are unlikely (Farrar & Glauber, 1967). For the regression analyses, standardized coefficients (β) are shown in Table 3.

For predictions at T_2, the independent variables (predictions at T_1, postdictions at $T_{1,}$ and performance at T_2) explained 28% of variance, adjusted $R^2 = 0.28$, $F(6, 267) = 18.81$, $p < .001$ As shown in Table 3, predictions and postdictions at T_1 affected predictions at T_2, whereas actual performance at T_2 was not related to T_2 predictions. No interaction effects were significant.

For postdictions, the independent variables and interaction terms in the regression model explained 35% of variance, adjusted $R^2 = 0.35$, $F(6, 267) = 25.33$, $p < .001$. Although predictions at T_1 did not affect postdictions at T_2, the main effect of postdictions at T_1 shows that these judgments affected postdictions at T_2. Moreover, T_2 postdictions were positively related to concurrent T_2 performance$_.$ There were no significant interaction effects.

In sum, these analyses show that second and fourth-grade children's predictions at T_2 are most strongly related to the judgments they made one year earlier (both predictions and postdictions) rather than to their actual task performance. In contrast, postdictions made at T_2 are not only related to postdictions made at T_1, but also to children's actual task performance.

Discussion

The present longitudinal study investigated the accuracy and the stability of elementary school children's monitoring judgments over one year for two age groups (second and fourth graders) and two judgment types (predictions and postdictions). Children made global, task-specific monitoring judgments to subjectively evaluate their task performance. Accurate global judgments form a basis for successful self-regulation and efficient learning (Dunlosky & Rawson, 2012; Rinne & Mazzocco, 2014; Schneider & Löffler, 2016; Thiede et al., 2003).

Table 3
Regression coefficients indicating effects of monitoring judgments At T1 and actual performance on monitoring judgments *At T_2.*

	Predictions T_2		Postdictions T_2	
Variable	β	SE *B*	β	SE *B*
Predictions T_1	0.22*	0.09	0.05	0.09
Postdictions T_1	0.31**	0.09	0.42***	0.09
Performance T_2	0.13	0.08	0.37***	0.07
Predictions T_1 x Age Group	−0.03	0.12	−0.05	0.12
Postdictions T_1 x Age Group	0.11	0.12	0.08	0.12
Performance T_2 x Age Group	−0.12	0.10	−0.07	0.10

Note. All predictors and outcome measures were transformed into *z*-scores to enable comparison between second and fourth graders. * $p \le .05$; ** $p \le .01$; *** $p \le .001$.

Accuracy of predictions and postdictions

For predictions, at both measurement points, the relation between judgments and performance was weak, indicating inaccurate self-monitoring. In contrast to Hypothesis 1 that monitoring would be more accurate for older than younger children, prediction accuracy was comparably low for second and fourth graders. Further, for both grades, there was no improvement in prediction accuracy over one year.

Research on item-by-item judgment accuracy finds age-related improvements in monitoring accuracy (Roebers et al., 2019; Schneider & Löffler, 2016). The few studies on the accuracy of children's task-specific predictions also suggest better monitoring accuracy for older elementary school children (Lipko et al., 2012; Shin et al., 2007). However, in these previous studies, children were asked to recall single pictures or objects. In the present study, children memorized associations. The differences in the task format could explain why we did not find age differences in prediction accuracy (Steiner, van Loon, Bayard, & Roebers, 2019; Thiede & Dunlosky, 1994).

In line with previous studies with children using a short period between measurements (Lipko et al., 2009; 2012; Shin et al., 2007), our longitudinal findings show that prediction accuracy did not increase over one year. For adults, some studies found that prediction accuracy improved (Connor et al., 1997; Dunlosky & Hertzog, 2000; Hertzog et al., 1994), whereas other studies did not show any improvement over time (e.g., Foster et al., 2017). These mixed findings may be due to the task format and the delay between measurements. Particularly, studies with word-pair recall tasks using multiple measurements within one day, seem to document prediction improvements (e.g., Connor et al., 1997; Dunlosky & Hertzog, 2000). In contrast, studies assessing monitoring accuracy for college exam performance with longer delays between measurements (1 week or more; e.g., Foster et al., 2017) do not find prediction improvements. It may be that improvements in prediction accuracy are more apparent for recall tasks (in other studies; Thiede & Dunlosky, 1994) than for recognition tasks (as in our study). Future research could further investigate factors affecting the improvement of prediction accuracy over time, and assess the effects of the delay between measurements, the learning task format, and the test format.

For postdictions, in line with Hypothesis 1, monitoring accuracy was higher for fourth than for second graders. This finding confirms previous studies investigating item-by-item postdictions (Ghetti et al., 2006; Roebers, 2002; Roebers et al., 2019), and shows that age differences are also found when investigating monitoring accuracy with task-specific global postdictions. The finding that postdiction accuracy improved for the second, but not the fourth graders may suggest that the children's ability to make accurate postdictions mainly develops between second and third grade.

To our best knowledge, the present study is the first to allow for a direct comparison between the accuracy of children's task-specific predictions and postdictions. Our expectation that postdictions would be more accurate than predictions (Hypothesis 2) was partially confirmed. At the first measurement point, postdictions were more accurate than predictions in the fourth but not in the second grade. At the second measurement point, postdictions were more accurate than predictions, independent of age.

When comparing the current findings with previous research investigating item-by-item monitoring judgments (Bayard et al., 2021; Dougherty, Scheck, Nelson, & Narens, 2005; Maki, Shields, Wheeler, & Zacchili, 2005; Robey, Dougherty, & Buttaccio, 2017; Roebers et al., 2019), patterns of results seem similar. Elementary school children's prospective judgments (i.e., judgments made before test-taking) are less accurate than retrospective judgments (i.e., judgments made after test-taking), and prospective monitoring judgments do not improve over time, regardless of whether these are made on a global or on an item-specific level. Moreover, our findings suggest that particularly for second graders, the accuracy of postdictions develops over time. Further, the ability to postdict performance appears to develop before the ability to predict. Presumably, the ability to accurately predict one's task performance is not yet developed in the fourth grade of elementary school (between 10 and 11 years of age), whereas the ability to accurately postdict seems to develop during the second grade school year (between eight and nine years of age).

Longitudinal monitoring judgment stability

One reason why monitoring judgments may not necessarily improve over time may be due to a tendency to make similar judgments, regardless of variances in learning experiences and test experiences. To further acquire insights into the effects of a habitual judgment tendency on children's self-monitoring, we investigated how judgments made one year earlier (i.e., at T_1) were related to current judgments (at T_2). Findings showed that the one-year stability of metacognitive judgments was moderate to high. For predictions for second and fourth graders, test-retest correlation values were 0.40 and 0.41, respectively, for postdictions, correlations were 0.49 for second graders and 0.52 for fourth graders. In line with our expectations (Hypothesis 3), this suggests that the rank-order of children's judgments did not change substantially over time.

For performance, stability was rather low for second graders and only low to moderate for fourth graders, implying that children's task performance varied more over time than their judgments did. Moreover, there was no significant difference between the stability of predictions and postdictions. That is, even though postdictions seemed more sensitive to variances in performance (i.e., postdictions more accurately reflected actual performance than predictions), these judgments were also based on habitual judgment tendencies. These findings imply that variable test experiences and stable judgment tendencies can jointly influence monitoring judgments.

The finding that judgments are related to each other over one year extends previous research showing short-term relations between judgments (Dunlosky & Hertzog, 2000; Hacker et al., 2000; Kleitman & Moscrop, 2010; Kleitman & Stankov, 2001, 2007; Lipko et al., 2012). The present study is the first to report children's longitudinal monitoring judgment stability and brings further evidence that individuals seem to have a habitual tendency to make judgments in a particular way. Interestingly, research on personality characteristics also investigated the annual stability of elementary school children's subjective judgments about their competencies, by assessing self-esteem and self-concept stability (Putnick et al., 2020; Trzesniewski et al., 2003). The stability of task-specific metacognitive monitoring judgments seems comparable to rank-order stability reported in research on global self-esteem (Trzesniewski et al., 2003), whereas annual stability of self-concept for scholastic competence (as reported by Putnick et al., 2020) seems much higher for elementary school children. This may indicate that particularly for self-concept, habitual tendencies explain almost solely why persons rate their competencies in a specific way; for metacognitive judgments, a global tendency explains judgment magnitudes to a lesser extent. Although metacognitive monitoring judgments are related to self-concept judgments (Dapp & Roebers, 2021), findings imply that besides rather stable person factors, also more variable task processes (for instance learning and test experiences, Efklides, 2011) affect metacognitive judgments.

A contribution of the present study is that we can offer insights into the joint effects of task factors and person factors on metacognitive monitoring judgments. When making predictions, children judged their performance right after learning. Interestingly, for both age groups, only one's previous judgments (both predictions and postdictions) influenced predictions one year later, whereas the predictions were not related to actual performance. This indicates that when predicting performance right after learning, children could not effectively implement their learning experiences to make accurate judgments, and mainly based their judgments on their habitual judgment tendencies. When making postdictions, children made judgments after obtaining test experiences. As for predictions, postdictions were strongly related to previous judgments, indicating effects of stable person factors on judgments. However, in contrast to predictions, postdictions were also related to actual performance, indicating that task factors related to self-test experiences benefitted judgment accuracy.

The cue utilization framework (Koriat, 1997) has frequently been used to theoretically ground research on metacognitive monitoring accuracy. According to this theory, individuals use a variety of available metacognitive cues (for instance processing fluency, ease of retrieval, or feelings of familiarity) to base their monitoring on. Cues are valid when these are related to performance, and when valid cues are utilized when self-monitoring learning, judgments will be accurate. The comparison between predictions and postdictions indicates that not the experiences made during learning Kanji associations, but rather the recognition test experiences provide children with valid cues. These findings confirm research with adult participants showing that self-testing can improve monitoring accuracy (Barenberg & Dutke, 2019; Händel, de Bruin, & Dresel, 2020). Through testing, participants can use the information from their recognition attempts as cues for their postdictive monitoring judgments. Findings may have implications for how teachers train children to self-monitor their learning in school. From third grade onwards, children seem able to account for test experiences when monitoring learning, asking them to complete a self-test before making judgments would seem beneficial.

However, even after accessing and using valid performance cues for postdictions, children's habitual judgment tendencies were not overruled. That is, in addition to being related to actual performance, the judgments that were made one year earlier were strongly related to postdictions. This implies that even after accessing and using valid cues for judgments, the influence of habitual judgment tendencies on self-monitoring remains.

Limitations and directions for future research

Although the present study shows how past judgments (predictions and postdictions), as well as actual task performance, affect children's self-monitoring judgments, at least 65% of the variance in judgments remained unexplained. Task materials, motivational factors, epistemological beliefs, and personality dispositions may all influence metacognitive judgments (Destan, Spiess, de Bruin, van Loon, & Roebers, 2017; Efklides, 2011; Händel et al., 2020). It is unclear how these factors simultaneously contribute to children's monitoring; future studies could further investigate this.

A strength of the used Kanji-task is that it is well validated to detect age differences and developmental progression in elementary school children's monitoring accuracy (Destan et al., 2014; Destan & Roebers, 2015; Roderer & Roebers, 2010; Roebers et al., 2019). However, although this task may be comparable to educational tasks such as vocabulary learning or fact learning, the task format may not be comparable to learning tasks for which deep comprehension and knowledge application are necessary. To address specific classroom implications, future studies could use other educational tasks when investigating children's monitoring skills. Further, monitoring is considered a basis for the regulation of learning (Dunlosky & Rawson, 2012; Schneider & Löffler, 2016). The design of our study does not allow concluding how judgments influenced children's strategy use and self-regulated learning activities; future research should address this issue.

For both age groups, monitoring judgments were higher at the last than the first measurement point. These increases in judgment magnitudes over one year may further emphasize similarities between research on self-monitoring and research on personality traits, as studies showed that values of children's mean evaluations of their competencies increase during the elementary school years (Orth et al., 2020). Further, performance was higher at the last measurement point. Even though the test items for both tests had a comparable difficulty level (as indicated by the item difficulty index; Moosbrugger & Kelava, 2008), it may have been easier for children to solve the tasks one year later, due to age-related cognitive development.

Moreover, children's repeated task experiences may have had positive effects on performance and judgments at the last measurement point (Bayard et al., 2021; Händel et al., 2020; Schraw & Roedel, 1994). Note that children were part of a larger study, for which they, during one year, also participated in five other measurements. For three measurements they learned texts, for one measurement they learned a secret code, and for one additional measurement (half a year in between the measurements used for the present study) they completed another Kanji learning task. Although task materials were new for each measurement, children became familiar with the procedure (i.e., learning, predicting performance, taking a test, and making postdictions). To investigate annual rank-order stability, for the present research, the first and last Kanji measurements were compared, but it needs to be taken into account that children had made experiences with the research procedure during the course of the year. Also in a natural school setting, children's learning experiences accumulate over the school years. The fact that children participated in multiple measurements does not seem to affect our conclusions about monitoring judgment accuracy and stability. However, future research could further investigate how repeated task experiences in research settings and school can affect judgments and performance.

Conclusions and implications

The present study adds new findings to the existing literature about factors influencing elementary school children's monitoring judgments. Particularly, findings bring insights into how the timing of the monitoring task can affect monitoring accuracy, and how monitoring judgments at a specific time point may affect future self-monitoring. Findings show that from third grade onwards, children's postdictions were more accurate than predictions. Moreover, this research shows that children's predictions and postdictions were stable over one year, indicating that judgments were influenced by habitual judgment tendencies. The present research can inform researchers and practitioners about how accurate children can judge their performance. In practice, findings may help teachers (and also parents and caretakers) to better understand when children struggle with monitoring their learning. Children who make high or low judgments appear to maintain their relative standing over a longer duration of time. That is, children may not sufficiently take variances in their actual learning into account when judging learning, but rather make similar judgments, regardless of changes in performance. Moreover, findings may help practitioners with the design of supportive learning environments to foster accurate self-monitoring. It seems that particularly before test-taking, making accurate judgments remains challenging. Children's postdictions, which were made after they completed the test questions and were presented with feedback about the correctness of their responses, were more accurate than their predictions. This may indicate that when aiming to improve children's judgments, using self-tests in combination with feedback on performance can support children with their self-monitoring. Possibly, giving children detailed feedback on their performance as well as on their monitoring accuracy could reduce reliance on habitual tendencies when self-monitoring learning, and support children to better implement their task-related study and test experiences. Future research could investigate how children's tendency to base judgments on a habitual response can be overruled, to make self-monitoring more accurate, and subsequent regulation more effective.

Author statement

The manuscript is original and does not include copyrighted materials.

The reported research was conducted according to the APA ethical standards.

Funding

This project was financed by a grant of the Swiss National Science Foundation (SNSF-Grant No. 100014_165477/1).

Availability of data and material

All materials and data can be obtained from the first author upon request.

Declaration of Competing Interest

None.

Acknowledgement

We like to thank the participating teachers and children for their cooperation, and the research assistants who helped us with the data collection. We further like to thank Jakob Raible and Stefan Kodzhabashev for their support with programming.

References

Barenberg, J., & Dutke, S. (2019). Testing and metacognition: Retrieval practice effects on metacognitive monitoring in learning from text. *Memory, 27*(3), 269–279. https://doi.org/10.1080/09658211.2018.1506481

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

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290. https://doi.org/10.3200/JEXE.73.4.269-290

Cole, D. A., Maxwell, S. E., Martin, J. M., Peeke, L. G., Seroczynski, A. D., Tram, J. M., & Maschman, T. (2001). The development of multiple domains of child and adolescent self-concept: A cohort sequential longitudinal design. *Child Development, 72*(6), 1723–1746. https://doi.org/10.1111/1467-8624.00375

Collins, A., Brown, J. S., & Newman, S. E. (1988). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children, 8*(1), 2–10. https://doi.org/10.5840/thinking19888129

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*(1), 50–71. https://doi.org/10.1037/0882-7974.12.1.50

Crone, E. A., & Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends in Cognitive Sciences, 21*(3), 205–215. https://doi.org/10.1016/j.tics.2017.01.003

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

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*(1), 188–200. https://doi.org/10.1037/0022-0663.95.1.188

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213–228. https://doi.org/10.1016/j.jecp.2014.04.001

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

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

Dignath, C., & Büttner, G. (2018). Teachers' direct and indirect promotion of self-regulated learning in primary and secondary school mathematics classes. Insights from video-based classroom observations and teacher interviews. *Metacognition and Learning, 13*(2), 127–157. https://doi.org/10.1007/s11409-018-9181-x

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

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*, 1096–1115. https://doi.org/10.3758/BF03193216

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*(3), 462–474. https://doi.org/10.1037/0882-7974.15.3.462

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

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky, & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–37). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.14.

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

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

Efklides, A., & Tsiora, A. (2002). Metacognitive experiences, self-concept, and self-regulation. *Psychologia, 45*(4), 222–236. https://doi.org/10.2117/psysoc.2002.222

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics, 49*(1), 92–107. https://doi.org/1937887.

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

Ghetti, S., Papini, S., & Angelini, L. (2006). The development of the memorability-based strategy: Insight from a training study. *Journal of Experimental Child Psychology, 94*, 206–228. https://doi.org/10.1016/j.jecp.2006.01.004

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test predictions and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Hacker, D. J., Bol, L., & Keener, C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 391–409). Psychology Press. https://doi.org/10.4324/9780203805503.

Händel, M., de Bruin, A. B. H., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning, 15*(1), 51–75. https://doi.org/10.1007/s11409-020-09220-0

Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1994). Metamemory and aging: Relations between predicted, actual and perceived memory task performance. *Aging and Cognition, 1*(3), 203–237. https://doi.org/10.1080/13825589408256577

Kail, R. (2015). *Children and their development* (7th ed.). New York: Pearson Academic.

Kleitman, S., & Moscrop, T. (2010). Self-confidence and academic achievements in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender. In A. Efklides, & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 293–326). New York: Springer. https://doi.org/10.1007/978-1-4419-6546-2_14.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*, 321–341. https://doi.org/10.1002/acp.705

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

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

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*, 265–279. https://doi.org/10.1016/j.jecp.2008.10.005

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 1–17). New York: Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4615-1099-4_1.

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning, 7*, 75–90. https://doi.org/10.1007/s11409-011-9079-3

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young children are not underconfident with practice: The benefit of ignoring a fallible memory heuristic. *Journal of Cognition and Development, 13*(2), 174–188. https://doi.org/10.1080/15248372.2011.577760

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

Lockl, K. (2010). Entwicklung des Metagedächtnisses bei Kindern und Jugendlichen [The development of metamemory in children and adolescents]. In H.-P. Trolldenier, W. Lenhard, & P. Marx (Eds.), *Brennpunkt der Gedächtnisforschung: Entwicklungs- und pädagogisch-psychologische Perspektiven* (pp. 191–212). Hogrefe.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchili, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731. https://doi.org/10.1037/0022-0663.97.4.723

Moosbrugger, H., & Kelava, A. (2008). *Testtheorie und Fragebogenkonstruktion* (p. 77). Springer-Verlag GmbH. https://doi.org/10.1007/978-3-662-61532-4

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

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2020). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000378

Putnick, D. L., Hahn, C. S., Hendricks, C., & Bornstein, M. H. (2020). Developmental stability of scholastic, social, athletic, and physical appearance self-concepts from preschool to early adulthood. *Journal of Child Psychology and Psychiatry, 61*(1), 95–103. https://doi.org/10.1111/jcpp.13107

Rinne, L. F., & Mazzocco, M. M. (2014). Knowing right from wrong in mental arithemtic judgments: Calibration of confidence predicts the development of accuracy. *PLoS One, 9*(7), Article e98663. https://doi.org/10.1371/journal.pone.0098663

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. (2017). Making retrospective confidence judgments improves learners' ability to decide what not to study. *Psychological Science, 28*, 1683–1693. https://doi.org/10.1177/0956797617718800

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

Roebers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38*(6), 1051–1067. https://doi.org/10.1037//0012-1649.38.6.1052

Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P. Bauer, & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (pp. 865–894). Wiley.

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

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

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring evidence from judgment-of-learning task. *Cognitive Development, 15*, 115–134. https://doi.org/10.1080/01650250143000210

Schraw, G., & Roedel, D. B. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22*(1), 63–69. https://doi.org/10.3758/BF03202762

Schunk, D. H., & Zimmerman, B. J. (Eds.). (2012). *Motivation and self-regulated learning: Theory, research, and applications*. Routledge.

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

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*, 747–758. https://doi.org/10.1016/j.lindif.2012.05.013

Steiner, M., van Loon, M. H., Bayard, N. S., & Roebers, C. M. (2019). Development of children's monitoring and control when learning from texts: Effects of age and test format. *Metacognition and Learning, 15*, 3–27. https://doi.org/10.1007/s11409-019-09208-5

Stipek, D. J., Roberts, T. A., & Sanborn, M. E. (1984). Preschool-age children's performance expectations for themselves and another child as a function of the incentive value of success and the salience of past performance. *Child Development, 55*(6), 1983–1989. https://doi.org/10.2307/1129773

Thiede, K., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*, 290–302. https://doi.org/10.1037/0022-0663.86.2.290

Thiede, K. W., Anderson, M. C., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology, 84*(1), 205. https://doi.org/10.1037/0022-3514.84.1.205

Van Loon, Mariette, H., Bayard, Natalie, S., Steiner, Martina, & Roebers, Claudia M (2021). Connecting teachers' classroom instructions with children's metacognition and learning in elementary school. *Metacognition and Learning, 16*(3), 623–650. https://doi.org/10.1007/s11409-020-09248-2

Van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology, 31*(5), 508–519. https://doi.org/10.1002/acp.3347