Journal of Experimental Child Psychology 241 (2024) 105857

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

The structure of metacognition in middle childhood: Evidence for a unitary metacognition-for-memory factor

Mariëtte van Loon[a,*], Ulrich Orth[b], Claudia Roebers[b]

[a] Department of Psychology, University of Zurich, 8050 Zurich, Switzerland
[b] Department of Psychology, University of Bern, 3012 Bern, Switzerland

ARTICLE INFO

Article history:
Received 25 May 2023
Revised 5 December 2023
Accepted 24 December 2023

Keywords:
Metacognition
Monitoring
Control
Learning
Development
Middle childhood

ABSTRACT

It has been debated whether children's metacognitive monitoring and control processes rely on a general resource or whether metacognitive processes are task specific. Moreover, findings about the extent to which metacognitive processes are related to first-order task performance are mixed. The current study aimed to uncover the relationships among children's monitoring (discrimination between correct and incorrect responses), control (accurate withdrawal of wrong answers), and performance across three memory-based learning tasks: Kanji learning, text comprehension, and secret code learning. All tasks consisted of a study phase, a test phase, monitoring (confidence judgments), and control (maintaining/withdrawing responses). Participants were 325 children (151 second graders [M_{age} = 8.12 years] and 174 fourth graders [M_{age} = 10.20 years]). Confirmatory factor analyses showed that a model in which monitoring and control loaded on a joint factor and performance on a separate factor provided the best fit to the data. Fourth graders had better monitoring and control accuracy than second graders. However, the factor structure of metacognition was similar for both age groups, contradictory to the assumption that metacognition generalizes across tasks as children grow older. After accounting for task-specific processes, monitoring and control skills for language-based memory tasks appear to be generalizable in middle childhood. In sum, children's monitoring

* Corresponding author.
E-mail address: mariette.vanloon@psychologie.uzh.ch (M. van Loon).

https://doi.org/10.1016/j.jecp.2023.105857
0022-0965/

and control for three separate memory tasks appear to reflect a unitary metacognition-for-memory factor related to, but distinguishable from, performance.

Introduction

The capacity to monitor and control learning, referred to as procedural metacognition, is a prerequisite for effective learning in and beyond the school context (Dunlosky & Rawson, 2012; Schneider & Löffler, 2016). Nelson and Narens (1990) explained procedural metacognition by proposing a "meta-level" structure of cognition through which monitoring and control processes operate to alter learning and memory (the "object level").

When monitoring, individuals judge how well they perform, that is, to what extent they have met learning standards. To obtain insight into monitoring accuracy, it has been recommended to elicit subjective judgments for each task item rather than assessing monitoring on a global level (e.g., judgments about exam grades; Dunlosky et al., 2016). Monitoring accuracy can be assessed with measures of discrimination (also referred to as relative accuracy or resolution) and measures of calibration (also referred to as absolute accuracy or bias). Discrimination measures are used to obtain insights into the differences between confidence for correct and incorrect responses. Calibration measures assess whether judgments align with objective performance and indicate to what extent individuals over- or underestimate their abilities. In this study, we focused specifically on discrimination measures with memory-based learning tasks. Discrimination is accurate when correct test responses are associated with higher confidence than incorrect answers (Roebers & Spiess, 2017). Accurate discrimination is a prerequisite for adequate control (Dunlosky & Rawson, 2012). When persons are uncertain that their responses are correct, this can be used as input to control learning activities (e.g., allocating restudy time to the items identified as not well learned). Furthermore, accurate discrimination improves the accuracy of memory reports because it allows a person to strategically submit responses that are held with high confidence and to withhold the answers for which the person is uncertain about correctness (Koriat & Goldsmith, 1996).

Development of monitoring and control

Children's monitoring ability starts developing in early childhood (Godfrey et al., 2023; Roebers, 2017; Schneider et al., 2022; Weil et al., 2013). For instance, when learning associations between images, 3-year-olds can differentiate between remembered and not-remembered items (Balcomb & Gerken, 2008). Furthermore, when judging whether an image is previously studied or new, 4-year-olds' correct responses are typically associated with higher confidence than incorrect responses (Hembacher & Ghetti, 2014). Skills to monitor learning appear to be related to the difficulty of the learning task and the format of the test questions (Steiner et al., 2020). The ability to monitor short recall test responses (e.g., paired-associate memory) appears to develop earlier than skills to monitor more complex learning (e.g., comprehension of causal relations in a text). The accuracy of recall monitoring appears to plateau around 10 years of age (Schneider & Löffler, 2016), whereas development of monitoring skills to judge complex perceptual decisions and text comprehension continues in adolescence (De Bruin et al., 2011; Weil et al., 2013).

Moreover, in early childhood control decisions seem to exhibit some degree of strategic consideration. Hembacher and Ghetti (2014) found that 3-year-olds more often withdrew answers when they were uncertain than when they were confident even though they were unable to monitor performance accurately. From 4 years of age, children are able to exclude incorrect responses more often than correct responses. Age-related differences in control are more pronounced for acting on incorrect responses compared with correct responses, such as allocating additional study time, searching additional information, and withdrawing errors from grading (Destan et al., 2014; Roebers & Spiess, 2017). Findings from studies investigating monitoring and control for educational tasks, such as learning vocabulary of a new language (Bayard et al., 2021), understanding information in videos (Roebers et al., 2014) and texts (Steiner et al., 2020), and learning spelling (Roebers & Spiess, 2017), tend to show that monitoring develops before control can be effectively implemented. Control skills show developmental improvement until late adolescence (Crone & Steinbeis, 2017; Roebers et al., 2014). Moreover, studies using educational tasks showed that accurate monitoring and effective control start to be beneficial for children's learning performance at 10 years of age (Rinne & Mazzocco, 2014; van Loon & Oeri, 2023). In particular, children's monitoring and control of incorrect performance is related to academic achievement (Selmeczy et al., 2021).

Interrelations among monitoring, control, and learning performance

Close relations between monitoring, control, and performance are theoretically assumed (Nelson & Narens, 1990). Monitoring is not directly predictive of learning but instead influences control decisions and actions, which then affect performance (De Bruin & van Gog, 2012; Dunlosky & Rawson, 2012; Nelson & Narens, 1990; Rinne & Mazzocco, 2014). Although monitoring and control are theoretically separable (Nelson & Narens, 1990), from 8 to 12 years of age these processes appear to be closely linked (van Loon & Roebers, 2017). For instance, van Loon et al. (2013) showed that the correlation between monitoring and restudy selections for 12-year-olds was as high as .93, and Steiner et al. (2020) reported correlations as high as .85 between judgments and maintenance/withdrawal decisions for 10-year-olds.

Despite close relations between monitoring and control, research with adults dissociated these processes. Peng and Tullis (2021) showed that only control, but not monitoring accuracy, was hindered when participants needed to divide their attention. Neuroimaging findings by Qiu et al. (2018) showed some overlap in monitoring and control, such that similar prefrontal cortex networks were activated. However, the dorsal anterior cingulate cortex was activated during uncertainty monitoring, whereas the inferior frontopolar cortex was activated during control. In addition, for children dissociations between monitoring and control were found; children diagnosed with autism spectrum disorder had lower monitoring accuracy but similar control skills than typically developing children (Grainger et al., 2016).

Moreover, metacognition and learning appear to be closely linked (Fleming & Lau, 2014; Vuorre & Metcalfe, 2022). It appears that superior performance in a task fundamentally changes the quality of involved metacognitive monitoring processes, such that higher performance is often associated with better monitoring and control (Roebers & Spiess, 2017; Vuorre & Metcalfe, 2022). When using discrimination measures to separately assess monitoring and control for correct and incorrect task performance, measures of metacognitive accuracy are less likely to be confounded with performance measures as when using calibration measures based on subtracting performance from judgments (Dunlosky et al., 2016). When using discrimination measures, monitoring and control remain to explain substantial amounts of individual differences in task outcomes (Roebers et al., 2014). The question arises whether this indicates that memory and metacognitive skills are so closely intertwined that they should be considered as an overarching higher-order capacity encompassing both cognitive and metacognitive skills. However, most evidence suggests that memory and metacognition are dissociable. For instance, adult patients with temporal lobe epilepsy showed impairments in memory but not in monitoring and control (Howard et al., 2010). Moreover, even when task performance is at chance level, participants may be able to discriminate between correct and incorrect responses (Scott et al., 2014). In the current study, we addressed to what extent elementary school children's monitoring, control, and task performance are dissociable when working on educationally relevant learning tasks.

Generalizability of metacognition across tasks

In education, children mainly work on tasks requiring memory (e.g., when learning a second language, when learning concepts) or comprehension (e.g., when studying text materials for subjects like history, biology, and geography). To better understand how metacognition can be trained and to what extent training effects would transfer, it would be highly relevant to obtain insights into whether children's metacognition should be considered a generalizable skill or whether this varies from task to task.

Adults' monitoring judgments across tasks appear to share a high degree of variance (Mazancieux et al., 2020; McCurdy et al., 2013). Because monitoring primarily relies on the prefrontal cortex, independent of task, monitoring is assumed to be generalizable (Fleming & Dolan, 2012). In contrast, Fitzgerald et al. (2017) found that adults' capacity to monitor across three tasks was dissociated, indicating that monitoring accuracy may also have task-specific elements. A meta-analysis by Rouault et al. (2018) showed that for adults, there were correlations between monitoring accuracy for different types of perceptual tasks. However, monitoring accuracy for perceptual tasks did not correlate with monitoring accuracy for memory tasks. Based on these findings, the authors suggested that a factor indicating metacognition-for-cognition may be distinguished from a metacognition-for-perception factor. This was further confirmed by a study by Lehmann et al. (2022), who found a unitary factor for metacognition for cognitive tasks unrelated to metacognition for perceptual discrimination tasks.

For children, limited research has addressed the generalizability of metacognition, given that most studies investigating children's procedural metacognition used only one task rather than multiple tasks. The few studies investigating the generalizability of children's monitoring skills indicate that this may be affected by developmental factors. For 5- to 8-year-old children, monitoring judgments for arithmetic and emotion discrimination tasks appeared to be unrelated (Vo et al., 2014). From 8 to 10 years of age, monitoring skills may become related across different tasks (Bellon et al., 2020; Geurten et al., 2018; Veenman & Spaans, 2005). Using arithmetic and spelling tasks, Bellon et al. (2020) found that for 7- and 8-year-olds monitoring accuracy was not correlated between tasks, whereas for 8- and 9-year-olds monitoring accuracy measures were related to each other. Geurten et al. (2018) asked children in three different age groups (8- and 9-year-olds, 10- and 11-year-olds, and 12- and 13-year-olds) to judge how confident they were to be able to use the best strategy for an arithmetic task and a word-pair-learning task. They found that metacognitive monitoring was not related across tasks for 8-year-olds but that from 10 years of age onward metacognition for the two different task domains became related. Furthermore, findings for 9- to 12-year-olds (Kleitman & Moscrop, 2010) and 12- to 14-year-olds indicate relations between children's monitoring across different tasks.

Together, this seems to indicate that in mid to late elementary school, children's metacognitive skills become generalizable across different tasks. Children's emerging ability to transfer their monitoring skills across tasks seems to be due to cognitive maturation as well as school practice and experiences with engaging in metacognition (Roebers et al., 2019). However, whether metacognition is generalizable or task specific has only been investigated for monitoring, and a limitation of previous research is that on-task measures within participants of metacognitive control have not been considered. With this research, we aimed to obtain insights into this issue by eliciting measures of not only monitoring but also control for three different language-based learning tasks for which information needed to be memorized.

The current study

This study aimed to provide unique insights into the structure of metacognition in middle childhood. Children completed three different learning tasks in the classroom. The three tasks used different types of learning and memory stimuli: a Japanese (Kanji) vocabulary learning task, a text comprehension task, and a secret code learning task matching symbols with letters. The tasks' phases were similar (i.e., learning, taking a test, monitoring test performance, and then controlling performance). However, the learning and test phases had different formats. That is, children completed (a) a paired-associated Kanji-learning task with fixed learning times and a four-alternative multiple-choice recognition test, (b) a text comprehension task with self-paced learning and a test with open-ended questions requiring words or short sentences as answered, and (c) a secret code memory task that was instructed by the classroom teachers and tested with a recall task requiring single-letter answers.

We describe age-related differences in monitoring and control accuracy between second graders (aged 7–9 years) and fourth graders (aged 9–12 years). We expected to find age differences in monitoring and control skills, with more pronounced age differences for text comprehension than for memory tasks. More important, beyond replicating previous findings on the development of monitoring and control, we aimed to address to what extent we found evidence for a general metacognitive ability in children, such that metacognitive monitoring and control skills are related across different tasks. If we found that procedural metacognition has generalizable components in childhood, this could imply that for children metacognition for learning processes can be considered a trait-like overarching factor.

Past research investigating relations between metacognitive accuracy across tasks mainly relied on correlational measures (Rouault et al., 2018). However, the focus on zero-order correlations has the limitation that, due to a task impurity problem, underlying commonalities may be masked (Miyake et al., 2000). That is, differences between tasks in non-metacognitive processing requirements (e.g., number processing vs. language processing, learning for recognition vs. comprehension, easy tasks vs. complex tasks) may mask underlying commonalities in metacognition. If these task-specific aspects are not controlled, the relations between the metacognition factors may be substantially biased. Recently, Lehmann et al. (2022) used confirmatory factor analyses (CFAs) to investigate latent constructs of metacognition and to test whether a domain-general monitoring ability could be identified across multiple paradigms. Using CFAs minimized the task impurity problem given that this approach statistically extracts commonalities across tasks involving metacognition and can measure this as a "purer" factor (Lehmann et al., 2022; Miyake et al., 2000). With a CFA approach, Lehmann et al. (2022) found that measures of monitoring accuracy for different memory tasks loaded on one latent factor (although metacognition for perceptual discrimination tasks loaded on another latent factor).

In the current research, using CFAs we compared potential structures of metacognition for the two age groups. Measures of monitoring accuracy, control accuracy, and learning performance for the three separate tasks were used to create latent factors. First, we analyzed whether monitoring, control, and performance are shared across tasks (i.e., indicate generalizable skills). Based on prior evidence from adult participants for memory-based tasks (Lehmann et al., 2022), we expected to find evidence for generalizable skills, particularly for monitoring accuracy. Second, we investigated to what extent metacognitive skills are (in)separable from learning performance. Third, we addressed to what extent monitoring and control are unitary or distinguishable skills. After identifying the best-fitting latent factor model, we compared second and fourth graders to investigate whether the model fit differs between age groups.

Two age groups (8-year-old second graders and 10-year-old fourth graders) were compared to determine whether age affects the structure of metacognition. Children in these age groups are in an important developmental period for metacognitive monitoring, control, and memory development. There are pronounced changes from 8 to 10 years of age in children's ability to use monitoring to guide control (Selmeczy & Ghetti, 2019; van Loon et al., 2013). Moreover, metacognitive abilities appear to shift from task-specific to general from 8 to 13 years of age (Bellon et al., 2020; Geurten et al., 2018). Therefore, we expected to find more robust evidence for generalizable metacognition-for-memory skills for the older age group (fourth graders, 9–12 years of age) than for the younger age group (second graders, 7–9 years of age).

Method

Participants and design

Participants (N = 325, 49.4% female) were 151 second graders (M_{age} = 8.1 years, SD = 0.49; 10 7-year-olds, 112 8-year-olds, and 29 9-year-olds) and 174 fourth graders (M_{age} = 10.2 years, SD = 0.50; 6 9-year-olds, 129 10-year-olds, 37 11-year-olds, and 2 12-year-olds). Written consent was obtained from their parents/caretakers. They participated in a large project that included seven measurements: in Fall Year 1 with Kanji and text tasks, in Spring Year 1 with Kanji, text, and secret code tasks, and in Fall Year 2 with Kanji and text tasks. Due to the scope of this research, the measurements in Spring Year 1, when participants completed all three tasks, were used. Findings for the Kanji task were reported in detail by Roebers et al. (2019), the text task was reported by Steiner et al. (2020), and the secret code task was reported by Van Loon et al. (2021). The current approach is the first to combine findings for all three tasks.

Materials and procedure

Children were tested in their school classes. During three sessions approximately 1 week apart, they completed a Kanji task, a text comprehension task, and a secret code task. All tasks consisted of learning, test, monitoring, and control phases. Fig. 1 shows these phases. All materials were previously piloted with a different sample.

Kanji task

On a tablet computer, children learned 12 (second graders) or 16 (fourth graders) Japanese characters (Kanjis, as used by Destan et al., 2014) and their meanings presented as pictures. Each Kanji and its meaning were randomly shown for 5 s when learning. For the recognition test, one Kanji at a time was presented with four alternatives (the correct alternative and three randomly selected alternatives that had appeared in the learning phase). There was no time limit to complete the test. After choosing an answer, a red frame surrounded it and the children made a monitoring judgment (CJ). They answered the question “How sure are you that you have chosen the correct picture?” by clicking on a 7-point thermometer scale ranging from *very unsure* to *very sure* (adapted from Koriat & Shitzer-Reichert, 2002). In the end, participants were presented with their answers, one at a time, and could either maintain or withdraw these by selecting a green or red traffic light, respectively. They were told that they could earn 1 point for maintaining correct answers but that 3 points would be deducted for maintaining incorrect answers (based on the +1, −3 bonus-to-penalty ratio by Roebers et al., 2009).

Text task

Six expository texts were read in randomized order on a tablet, with reading being self-paced. After reading, children received a booklet with 12 open-ended test questions (2 per text). Per text, 1 ques-

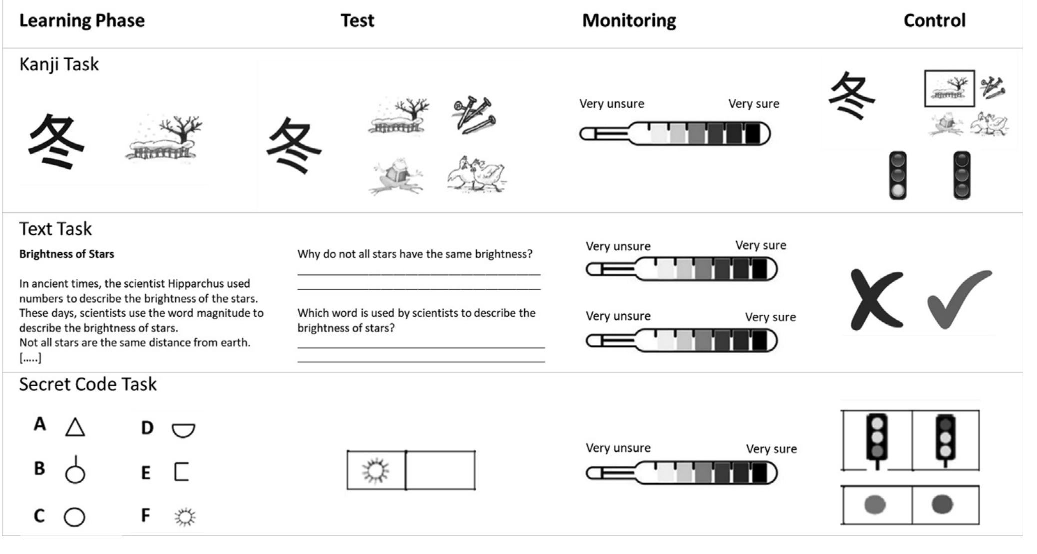

Fig. 1. Task phases for the three tasks.

tion required a sentence as an answer (comprehension question), and 1 required a single-word answer (detail question). For each answer, they then monitored how certain they were that their answer was correct by making a CJ on the 7-point thermometer printed to the right of each question. At the end, children decided for each of their answers whether they wanted to maintain it (by adding a check mark) or withdraw it (by crossing out the answer). They were told that the +1, −3 bonus-to-penalty ratio would be used to award or deduct points.

Secret code task

For this task, 26 alphabet letters were matched to 26 symbols (some of these pairs are depicted in Fig. 1). The teachers instructed the secret code task in the classroom. They received teaching materials from the researchers: a poster with the secret code and a practice booklet with four secret code reading tasks and five writing tasks. This way, all teachers could instruct the task using the same materials. However, teachers could decide how they used these materials for instruction. Teachers were also instructed that the children would be tested with a recall test for which they would need to write down the letters corresponding to the secret code symbols. Teachers could use 30 min for instruction and practice. At the beginning of the lesson, a timer was set. After 30 min when the timer went off, a researcher took over and the test booklets were handed out. For the recall test, children saw the symbols and needed to write down the matching letter. Next to each letter answer was printed the 7-point thermometer, on which they could indicate with a CJ how sure they were that their answer was correct. After completing all questions, children maintained or withdrew each of their answers using the red or green traffic light. They were informed about the +1, −3 bonus-to-penalty ratio.

Analyses

For each task, the percentage of correct test responses was calculated. For monitoring accuracy, for each child a difference measure was calculated between CJs for correct and incorrect task responses (as done by Destan & Roebers, 2015). A larger difference score indicates stronger discrimination (i.e., more accurate monitoring) between correct and incorrect responses. For control, the percentage of adaptive control decisions (incorrect answers withdrawn) was calculated per child (following Roebers & Spiess, 2017). First, we address whether we found differences between the age groups in performance, monitoring accuracy, and control accuracy; independent *t* tests were conducted to compare the mean values for the second and fourth graders.

Lavaan with R was then used to conduct a series of four CFAs, as depicted in Fig. 2. Because the values for performance and monitoring were left-skewed for the secret code task, for the CFAs these values were reflected and transformed with a reciprocal $x \rightarrow 1/x$ transformation (Cox, 2007). Furthermore, all performance, monitoring, and control scores were converted to *z* scores. By comparing the four models, we investigated whether performance, monitoring, and control are best represented by three latent constructs (Models 1 and 2), one unitary construct (Model 3), or two separate constructs (Model 4). All models used the same nine individual, *z*-transformed indicators: performance, monitoring, and control for the three tasks. To assess the fit of the hypothesized models, the χ^2 goodness-of-fit statistic, comparative fit index (CFI), and root mean square error of approximation (RMSEA) were used. Multigroup analyses were then computed for each of the four models by comparing one model for which factor loadings were constrained across age groups with a second model for which loadings were allowed to differ. This way, we could examine whether the loadings were invariant across the two age groups or whether there would be age-related differences. Finally, the best-fitting model was identified based on the χ^2 goodness-of-fit statistic and comparison of the Akaike information criterion (AIC) values using the R package "AICcmodavg" (Mazerolle, 2023).

Transparency and openness

The ethical committee of the Faculty of Human Sciences, University of Bern approved the materials and procedure. The design was not preregistered. Because we did not have clear indications of potential effect sizes, we could not conduct a priori power analyses to determine the sample size. However,

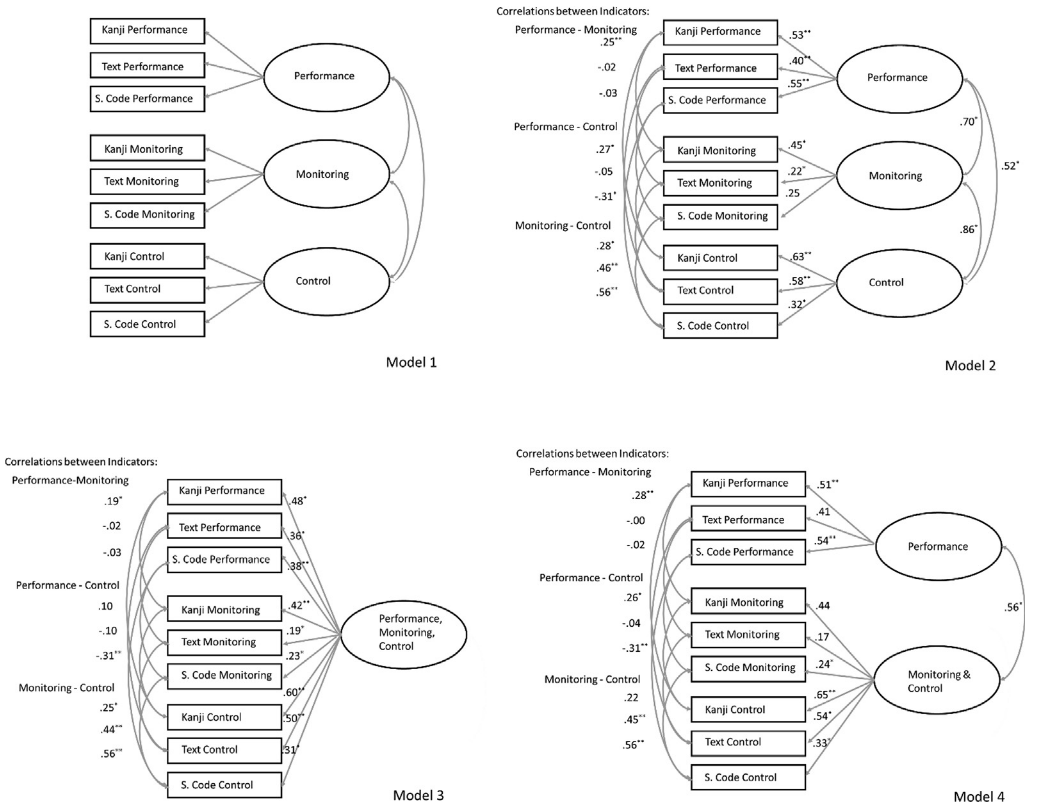

Fig. 2. Alternative models of the structure of metacognition. Significance is indicated with $^{*}p < .05$ and $^{**}p < .001$. For Model 1, factor loadings and correlations are not indicated because this model had a poor fit. S. Code, secret code.

to detect medium effect sizes with sufficient power, a sample of approximately 150 participants per age group would likely be appropriate (in line with Lehmann et al., 2022).

Data were analyzed with R and RStudio (Version 1.4.1717) and the R packages "lavaan" (Rosseel, 2012) and "AICcmodavg" (Mazerolle, 2023). The data and analysis code are available on the Open Science Framework (https://osf.io/da3jt/).

Results

Differences between age groups in metacognition and performance

Descriptive statistics for performance, monitoring, and control are presented in Table 1. Independent t tests between the age groups showed that for the Kanji task the fourth graders had higher task performance, $t(317) = 5.21$, $p < .001$, Cohen's $d = .58$, and monitored more accurately, $t(302) = 2.54$, $p = .012$, Cohen's $d = .29$, than the second graders. Although fourth graders tended to withdraw more incorrect Kanji responses than second graders, the difference in control accuracy was not significant ($p = .089$). For the text task, fourth graders had higher monitoring accuracy, $t(307) = 5.33$, $p < .001$, Cohen's $d = .61$, and had higher control accuracy, $t(310) = 4.70$, $p < .001$, Cohen's $d = .53$, than second graders. There were no significant differences between age groups in performance ($p = .854$) for the text task. For the secret code task, there were differences between age groups in performance, $t(303) = 3.60$, $p < .001$, Cohen's $d = .41$, but there were no age differences in monitoring accuracy ($p = .957$) and control accuracy ($p = .367$).

Table 1
Performance, monitoring, and control in both age groups

Task	Performance	Monitoring judgments (magnitudes)	Control (maintained answers)	Monitoring accuracy	Control accuracy
Second graders					
Kanji	53.64% (20.87)	4.72 (1.17)	54.40% (23.63)	1.24 (1.23)	64.64% (31.65)
Text	57.29% (22.10)	4.82 (1.28)	64.98% (22.60)	1.18 (1.48)	50.74% (33.13)
Secret code	84.62% (15.42)	6.16 (1.01)	88.13% (14.02)	2.91 (2.20)	63.86% (39.35)
Fourth graders					
Kanji	65.61% (20.11)	5.13 (1.18)	59.13% (22.18)	1.60 (1.23)	70.80% (31.18)
Text	57.75% (22.59)	4.56 (1.19)	58.09% (20.71)	2.03 (1.34)	67.52% (29.93)
Secret code	90.44% (12.84)	6.47 (8.81)	91.01% (13.97)	2.90 (2.16)	61.87% (41.03)

Note. Shown are mean values for performance (% correct responses), monitoring judgments (CJs; magnitudes ranging from 1 to 7), control decisions (% responses maintained), monitoring accuracy (discrimination of CJs between correct and incorrect answers), and control accuracy (% withdrawn incorrect responses). Standard deviations are in parentheses.

Correlations

Table 2 shows the correlations between performance, monitoring accuracy, and control accuracy for the three tasks for the two age groups; correlations for the second graders are presented above the diagonal, and findings for the fourth graders are presented below the diagonal. For the Kanji task, task-specific (i.e., within-task) correlations between performance and monitoring, between performance and control, and between monitoring and control appeared to be significant for both age groups. For the text task, for both age groups the correlations between performance and monitoring and between performance and control were low and nonsignificant, whereas there were significant correlations between monitoring and control. For the secret code task, for both age groups there were no significant correlations between performance and monitoring and between performance and control; the correlations between monitoring and control were significant.

Across-task correlations of performance were not significant between the Kanji and text tasks for second graders; this correlation was significant for fourth graders. Furthermore, for both age groups there were significant correlations between performance on the Kanji and secret code tasks and between performance on the text and secret code tasks. For monitoring accuracy, there were no significant correlations between tasks for both age groups. For control accuracy, correlations between the Kanji and text tasks and between the Kanji and secret code tasks were significant for both age groups; correlations between the text and secret code tasks were significant for second graders but not for fourth graders.

Table 2
Correlations between indicators of performance, monitoring accuracy, and control accuracy for both age groups

Variable	1	2	3	4	5	6	7	8	9
1. Performance Kanji	–	.098	.297**	.224**	.168*	.092	.291**	.184*	.075
2. Performance text	.256**	–	.252**	.155	.029	.120	.271**	.056	.056
3. Performance secret code	.201**	.266**	–	.037	.066	.076	.185*	.118	−.113
4. Monitoring Kanji	.418**	.129	.165*	–	.150	.124	.448**	.260**	.246*
5. Monitoring text	−.098	.081	−.006	.052	–	.159	.060	.418**	.075
6. Monitoring secret code	.135	−.083	.014	−.016	−.045	–	.132	.156	.530**
7. Control Kanji	.360**	.107	.044	.383**	−.026	.181	–	.379**	.232*
8. Control text	.050	.124	.052	.167*	.432**	.152	.316**	–	.241*
9. Control secret code	.222*	−.004	−.233	.032	.028	.651**	.223*	.134	–

Note. Correlations are presented above the diagonal for second graders and below the diagonal for fourth graders.
* $p < .05$.
** $p < .01$.

The structure of procedural metacognition

We created four models to address our research questions. These models, including information about factor loadings and correlations, are shown in Fig. 2.

First, to test to what extent children's procedural metacognition can be considered a general construct or whether it also has task-specific elements, we compared Model 1 with Model 2. In these models, the latent constructs (i.e., performance, monitoring accuracy, and control accuracy) were measured by three indicators each. In Model 1 (the general three-factor model), the task-specific indicators were not allowed to covary.

As shown in Table 3, Model 1 had a poor fit. This indicated that metacognitive skills could not be considered purely generalizable across tasks. Furthermore, there was no significant difference between the constrained and nonconstrained Model 1 ($\Delta\chi^2$ = 7.63, Δdf = 6, p = .27), indicating that the loadings did not differ significantly between the age groups. The lower AIC value indicated that the constrained model would be preferable (AIC constrained model = 7154.36, AIC nonconstrained model = 7158.73; ΔAIC = 4.37).

When residual correlations were included between indicators per task to account for task-specific differences (Model 2), this model resulted in a good fit (see Table 3). Compared with Model 1, the fit of Model 2 was significantly better ($\Delta\chi^2$ = 151.81, Δdf = 9, p < .001; ΔAIC = 133.81). This indicates that the components of metacognition are generalizable across tasks, but only when accounting for task-specific variance. Furthermore, there was no significant difference between the constrained and nonconstrained Model 2 ($\Delta\chi^2$ = 2.96, Δdf = 6, p = .81), indicating that the loadings held across the age groups. The constrained model appeared to be preferable (AIC constrained model = 7031.11, AIC nonconstrained model = 7040.15; ΔAIC = 9.04).

Moreover, with Model 3 (Fig. 2), we investigated to what extent performance, monitoring, and control can be considered independent constructs or whether these factors reflect a unitary construct. In this model, all indicator variables loaded on the same latent factor (reflecting a general capacity consisting of metacognitive accuracy and cognitive performance). To account for task-specific variance, indicators of the same task were allowed to covary. As shown in Table 3, this model had a good fit. There were no differences between the constrained and nonconstrained Model 3 ($\Delta\chi^2$ = 5.84, Δdf = 8, p = .67), indicating that the loadings held across age groups; the constrained model would be preferable (AIC constrained model = 7039.43, AIC nonconstrained model = 7049.59; ΔAIC = 10.1 6). However, Model 2 did fit the data better than Model 3 ($\Delta\chi^2$ = 19.48, Δdf = 3, p < .001), as indicated by higher CFI, lower RMSEA, and lower AIC scores (ΔAIC = 13.48). This suggests that separating monitoring, control, and performance better reflects the data than combining these in a unitary construct.

Furthermore, we investigated with Model 4 (Fig. 2) whether there is evidence that monitoring and control reflect a general metacognitive factor that can be separated from performance. All indicators of monitoring and control loaded on a latent metacognition factor, and the performance indicators loaded on a separate task performance factor. Indicators of the same task were allowed to covary. As shown in Table 3, this model provided a good fit for the data. Furthermore, for Model 4 there was no significant difference between the constrained and nonconstrained models ($\Delta\chi^2$ = 1.89, Δdf = 7; p = .97). Moreover, the lower AIC value indicated that the constrained model is preferable (AIC constrained model = 7023.3, AIC nonconstrained model = 7035.4; ΔAIC = 12.11). The fit for Model 4 was significantly better than the fit for Model 3 ($\Delta\chi^2$ = 16.83, Δdf = 1; p < .001). Although Model 2 and Model 4 did not significantly differ in fit ($\Delta\chi^2$ = 2.64, Δdf = 2; p = .27), the AIC index for Model 4 was lower than that for Model 2 (ΔAIC = 1.36), indicating that it would be preferable to consider mon-

Table 3
Fit measures for the four alternative models of the structure of metacognition

Model	*Df*	Chi-square	CFI	RMSEA	AIC
1	24	162.03	.647	.133	7189.36
2	15	10.22	1.000	.000	7055.55
3	18	29.70	.970	.045	7069.02
4	17	12.86	1.000	.000	7054.19

Note. CFI, comparative fit index; RMSEA, root mean square error of approximation; AIC, Akaike information criterion.

itoring and control a general metacognitive skill instead of being two different types of skills. In Model 4, the correlation between the latent metacognition and performance factors was significant, supporting the notion that, despite being distinguishable, metacognition and performance are related.

In sum, the results suggested that metacognitive accuracy has generalizable and task-specific elements given that a model only accounting for generalizable factors (Model 1) did not fit the data. Model 2 (the three-factor model), Model 3 (the unitary model), and Model 4 (the two-factor model) all fitted the data, and Model 2 and Model 4 fitted significantly better than Model 3. Based on inspection of the AIC fit indices and in accordance with the principle of parsimony, Model 4 was selected as the preferable model. As noted above, in all models the factor loadings were invariant across age groups. Thus, for the second- and fourth-grade children, monitoring, control and task performance across memory tasks seem to be best reflected by a unitary metacognition factor (comprising both monitoring and control) and a separate performance factor.

Discussion

The current study addressed the factor structure of children's metacognition to investigate (a) whether procedural metacognitive processes are generalizable across tasks and (b) whether monitoring, control, and performance are separable constructs. For second- and fourth-grade elementary school children, monitoring accuracy, control accuracy, and performance were investigated when completing Kanji recognition, text comprehension, and secret code recall tasks.

When investigating monitoring and control for the separate tasks, our findings about development of procedural metacognition aligned partially with previous research on age-related metacognitive development. The literature suggests that monitoring skills precede control skills given that effective control relies on accurate monitoring (Schneider & Löffler, 2016; Schneider et al., 2022). However, our findings on monitoring and control for three different memory tasks revealed a more nuanced picture and suggest that developmental differences in monitoring and control may depend on the task type.

For the Kanji task, age differences were mainly apparent for monitoring accuracy; although older children seemed to withdraw incorrect responses more adaptively, the difference between second- and fourth-grade children was not significant. For the text task, both monitoring and control skills appeared to develop; fourth graders outperformed second graders when discriminating between correct and incorrect responses and when withdrawing erroneous answers. Notably, there were no age-related differences in metacognitive monitoring and control for the secret code task despite higher task performance for the older age group.

Although all tasks investigated metacognition for language-based learning tasks, differences between the task materials and the performance tests may explain why age-related patterns are not uniform. For the Kanji task, it appears that development of control skills might have plateaued earlier than that of monitoring skills. This extends previous work by Hembacher and Ghetti (2014), who found that 3-year-olds were able to control without being able to monitor accurately, and suggests that also for older children in some cases control may precede monitoring. Monitoring Kanji recognition on a multiple-choice test, which requires accounting for the chance of guessing a correct answer, may have been challenging even for 10-year-old children (De Carvalho Filho, 2009). It is possible that deciding whether to submit or withdraw responses was easier for them than making fine-grained judgments about their (un)certainty that an answer would be correct. This could explain why, for the Kanji task, the development of monitoring did not appear to precede the development of control.

Monitoring text comprehension appeared more challenging than monitoring recall (De Bruin et al., 2011; Thiede & Dunlosky, 1994). This may explain our findings that monitoring and control showed developmental improvement for the text task but not for the secret code task. For the secret code task, it may have been easier for children to monitor response correctness and to withdraw answers when they lacked certainty. Even 8-year-olds were proficient in monitoring and control accuracy for secret code recall, and development of metacognition appeared to stabilize despite age-related changes in task performance.

By using CFAs to understand the structure of children's metacognition, we aimed to go beyond describing developmental differences in metacognition for different types of learning tasks. To our knowledge, this study is the first to address variation in children's metacognitive monitoring and control accuracy across tasks. Findings indicate that children's metacognitive skills for the three memory- and language-based learning tasks appear to be generalizable; measures of monitoring accuracy for the three different tasks loaded on one latent monitoring factor. This may imply that in middle childhood monitoring skills are, to some extent, generalizable (confirming findings by Veenman & Spaans, 2005, and Vo et al., 2014). However, the latent factor model fitted the data well only when accounting for task specificity. This supports the idea that monitoring accuracy also has task-specific elements (Fitzgerald et al., 2017).

A unique contribution is that metacognition was investigated with not only monitoring but also control measures. Accurate monitoring supports learning only when judgments are translated into accurate control actions (De Bruin & van Gog, 2012). The current research is the first to show that children's control skills are generalizable across the different memory tasks. Furthermore, children's monitoring and control accuracy were closely linked, and the model for which these monitoring and control indicators all loaded on one unitary metacognition factor provided the best fit for the data. The shared variance across monitoring and control suggests the involvement of a shared resource.

It has been suggested that metacognition is mainly task specific for young children and then generalizes from 8 to 13 years of age (Bellon et al., 2020; Geurten et al., 2018; Veenman & Spaans, 2005). Surprisingly, we found that metacognition for the 8-year-olds was more generalizable than assumed in previous studies; comparisons between the second and fourth graders did not show differences in the factor loadings. Even for children as young as 8 years, monitoring and control appeared to be best reflected by one unitary metacognition-for-learning factor. This generalizability for both age groups was found despite developmental differences between the age groups in monitoring and control. Although the younger age group showed less accurate monitoring for Kanji and less accurate monitoring and control for text than the older children, monitoring and control skills seemed to generalize across these memory-based learning tasks regardless of age. That is, when being able to accurately monitor and adaptively control for one task, it appeared likely that these skills would transfer to other tasks for younger and older children alike. Thus, it appears that developmental differences in the absolute level of monitoring and control did not affect the extent to which metacognitive processes appeared to be generalizable.

One reason why generalizability of metacognition appeared at an earlier age than in prior research may be because the tasks used in the current research were more similar to each other than those used in previous studies. Prior research combined very different tasks (e.g., arithmetic calculations, spelling, learning of word combinations, emotion discrimination; Bellon et al., 2020; Geurten et al., 2018; Vo et al., 2014). For this study, all tasks were language-based learning tasks administered in an educational setting. Task characteristics differed (e.g., self-paced learning for the texts, fixed learning times for Kanji, teacher instructions for the secret code task), and test formats were different (recognition for the Kanji, open-ended questions for the texts, and recall for the secret code). However, each task asked children to memorize information and followed the same procedure (study, test taking, monitoring with the use of a 7-point scale, and control by making binary maintenance/withdrawal decisions). It is crucial to recognize that the shared methodological aspects, language-based nature of the tasks, and memory components across tasks might have contributed to the observed factor structure; this may explain why we found that metacognition is generalizable at an earlier point in development than suggested by previous studies. Moreover, the sample size in the current research was larger than the sample sizes in most of the earlier studies investigating the task specificity and domain generality of children's metacognition (a maximum of 24 participants per age group in Geurten et al., 2018, and a maximum of 18 participants per age group in Vo et al., 2014, but note that Bellon et al., 2020, had a sample size of 147 participants per age group). With larger sample sizes, analyses may be more sensitive and find indications of generalizability even for younger age groups.

A further reason why we found evidence for the generalizability of metacognition across memory tasks for both age groups may be using a CFA approach rather than relying on the inspection of zero-order correlations between measures. When basing conclusions only on the pattern of correlations, it would appear that between-task correlations were generally low for monitoring (.15 or lower) and low to moderate for control (.32 or lower). Low zero-order correlations between measures correspond with results by Miyake et al. (2000), showing low correlations between tasks measuring executive functions. Lehmann et al. (2022) also found low correlations between monitoring accuracy for different memory tasks even though a latent factor for memory monitoring was found. These low correlations indicate that it is important to assume that metacognitive accuracy is affected by task-specific processes but cannot be interpreted as evidence that metacognition has no generalizable processes. A weakness of only relying on correlations is that it does not become clear whether low correlations are a reflection of independence of measures because differences in non-metacognitive task processing requirements (e.g., differences in the type of learning materials, task difficulty, or test processes) may mask the existence of commonalities between measures. The CFA approach minimizes this task impurity problem and provides a stronger assessment of the relationships between cognitive and metacognitive processes (Lehmann et al., 2022; Miyake et al., 2000).

The CFAs showed that the models accounting for a combination of unitary and task-specific processes in metacognition most accurately represent the data. The current data do not bring direct insights into the mechanisms underlying this found unitary factor for monitoring and control. Monitoring and control are not presumed to be entirely similar processes; CJs are inferential, such that these are based on specific cues (e.g., the fluency with which information came to mind, feelings of familiarity; Koriat, 1997). Children as young as 9 years appear to use fluency cues for their confidence judgments (van Loon et al., 2017). Control processes may be more strongly influenced by attention, working memory capacity, and intrinsic motivation than CJs (Peng & Tullis, 2021; Qiu et al., 2018). However, despite these differences in monitoring and control, our findings indicate that these processes rely on a shared resource. Mazancieux et al. (2020) interpreted evidence for the generalizability of metacognitive monitoring as a potential "*g*-factor" for metacognition. Our research findings may indicate that such a unitary metacognition *g*-factor is also visible in children, at least when using memory tasks, and that this factor consists of both monitoring and control. Because the term *g*-factor has mainly been used in intelligence research, the question arises as to whether metacognitive skills are an indicator of intelligence. For measures of metacognition for memory tasks, Ohtani and Hisasaka (2018) showed moderate correlations between metacognitive accuracy and intelligence. However, metacognition more strongly predicted achievement and remained a predictor of academic achievement after controlling for intelligence. This implies that metacognition and intelligence are at least partially separable and that the found unitary factor is unlikely to reflect intelligence per se. In addition, research on executive functions found that different tasks appear to reflect a unitary factor (Miyake et al., 2000). This may suggest that, more generally, unitary constructs are found when assessing the structure of higher-order cognitive functions.

A major limitation of the current research is that all used tasks were language-based learning tasks that belong to the memory domain. Although our findings bring evidence for generalization of metacognitive skills across these tasks, the current study cannot address to what extent metacognition is domain general when other types of tasks are used (e.g., when combining memory tasks with arithmetic problem-solving tasks and perceptual decision-making tasks). Further research should address the extent to which metacognitive monitoring and control accuracy generalize across tasks from different domains.

It is important to note that our measurement of monitoring and control may have affected our findings. For each test question answered, children provided confidence judgments during the monitoring phase and then decided in the control phase whether to keep or withdraw their responses. For all three tasks, the moderate to strong correlations between monitoring and control indicate that accurate monitoring influences the accuracy of subsequent control decisions. It must be noted that even though control decisions were made several minutes after CJs, the act of monitoring before controlling might have influenced the latter. Conclusions that monitoring and control share a common factor go beyond the examination of task-specific associations and are based on the connections between the latent monitoring and control constructs. Nonetheless, future research should aim to replicate these findings with a clearer separation between CJs and control decisions. For example, using different sets of items for monitoring and control or changing the order of the monitoring and control phases rather than consistently starting with monitoring may provide further insights into the similarity and separability of these metacognitive processes.

Even though the current research showed that it is possible to distinguish between latent factors of task performance and metacognition, the two factors were strongly correlated. Research has been criticized for not sufficiently dissociating between measures of performance and metacognition (Fleming & Lau, 2014). Particularly when global confidence judgments (i.e., single predictions or postdictions about performance) are used and when monitoring is assessed with calibration measures (i.e., indicating over- or underconfidence), confounds between performance and metacognition appear to be likely (Dunlosky et al., 2016). The current study aimed to circumvent these issues by using fine-grained trial-by-trial indicators of monitoring and control and discrimination measures to separate judgments for correct and incorrect responses (Dunlosky et al., 2016; Roebers et al., 2014). For the text and secret code tasks, performance was not significantly correlated with monitoring and control; that is, a confound does not appear to be likely. However, relations between performance and metacognition for the Kanji task appeared to be low to moderate but significant. This may reflect findings by Vuorre and Metcalfe (2022) that, particularly when using recognition tests, it appears to be challenging to separate performance and metacognition.

Recently, measures based on signal detection theory, such as meta-*d'*/*d'* and hierarchical (H) meta-*d'*, were developed to investigate monitoring accuracy while controlling for individual differences in the use of the confidence scales and task performance (Fleming, 2017; Rouault et al., 2018). However, these measures appear to be mainly useful for tasks with many test and monitoring trials (Rouault et al., 2018) when binary scales are used to assess confidence (but for an exception, see Lehmann et al., 2022) and when difficulty levels and performance scores are kept constant. To the best of our knowledge, these analytical approaches have not yet been applied to investigate children's metacognition in educational settings, where performance differences and a relatively low number of test questions and confidence ratings are the norm. Future research should pay further attention to the relations between performance and metacognition and could address whether separating these measures may ensure "purer" measurement of metacognition. At the same time, it is also important to realize that in educational settings, rather than considering relations between performance and metacognition only as a confound, it may be crucial to better understand the interplay between metacognition and performance. Previous studies indicate that task performance influences monitoring (Roebers & Spiess, 2017; Vuorre & Metcalfe, 2022). That is, when persons are not well able to memorize the learned materials, they are likely to be unsure, independent of whether answers are correct. In contrast, confidence is likely to be high when a person experiences fluent retrieval of most of the studied information. When metacognition is influenced by performance and children with lower performance levels exhibit reduced abilities in discrimination and control, these insights have implications for understanding metacognitive processes in educational contexts.

Conclusions

For this research, the structure of metacognition was investigated in a sample of children rather than a typical sample of young college students. A novelty of the current approach is that, in addition to monitoring, control measures were used as indicators of metacognition. Our findings are particularly relevant for understanding the generalizability of children's metacognition for learning when working on memory and text comprehension tasks in school. Metacognitive monitoring and control skills for these tasks had a degree of unity in children as young as 8 years. This indicates that, despite task-related developmental differences, monitoring and control constructs develop and operate in conjunction as early as middle childhood. Task-specific and general cognitive resources seem to be involved in children's metacognition. These findings may lead to promising directions because this may indicate that when children are trained to monitor and control their learning for a specific memory-based task, they may transfer these skills to other memory tasks. Future research could investigate to what extent results extend to more diverse domains such as mathematics and perceptual decision-making tasks.

CRediT authorship contribution statement

Mariëtte van Loon: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Ulrich Orth:**

Formal analysis, Writing – review & editing. **Claudia Roebers:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – review & editing.

Data availability

The data are openly available on OSF through the link https://osf.io/da3jt/.

Acknowledgment

This research was supported by the Swiss National Science Foundation (100014_165477/1).

Data availability

The data and analysis code are openly available on the Open Science Framework (https://osf.io/da3jt/).

References

Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science, 11*(5), 750–760. https://doi.org/10.1111/j.1467-7687.2008.00725.x.
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.
Bellon, E., Fias, W., & De Smedt, B. (2020). Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLoS One, 15*(3), e229932. https://doi.org/10.1371/journal.pone.0229932.
Cox, N. (2007). *Transformations: An introduction.* Retrieved December 6, 2022, from http://fmwww.bc.edu/repec/bocode/t/transint.html.
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.
De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005.
De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003.
De Carvalho Filho, M. K. (2009). Confidence judgments in real classroom settings: Monitoring performance in different types of tests. *International Journal of Psychology, 44*(2), 93–108. https://doi.org/10.1080/00207590701436744.
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*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z.
Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. U. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–37). Oxford University Press.
Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.
Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition, 49*, 264–277. https://doi.org/10.1016/j.concog.2017.01.011.
Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness, 2017*(1), nix007. https://doi.org/10.1093/nc/nix007.
Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417.
Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Hunan Neuroscience, 8*, 443. https://doi.org/10.3389/fnhum.2014.00443.
Geurten, M., Meulemans, T., & Lemaire, P. (2018). From domain-specific to domain-general? The developmental path of metacognition for strategy selection. *Cognitive Development, 48*, 62–81. https://doi.org/10.1016/j.cogdev.2018.08.002.
Godfrey, M., Casnar, C., Stolz, E., Ailion, A., Moore, T., & Gioia, G. (2023). A review of procedural and declarative metamemory development across childhood. *Child Neuropsychology, 29*(2), 183–212. https://doi.org/10.1080/09297049.2022.2055751.
Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Consciousness and Cognition, 42*, 65–74. https://doi.org/10.1016/j.concog.2016.03.003.
Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768–1776. https://doi.org/10.1177/0956797614542273.

Howard, C. E., Andrés, P., Broks, P., Noad, R., Sadler, M., Coker, D., & Mazzoni, G. (2010). Memory, metamemory and their dissociation in temporal lobe epilepsy. *Neuropsychologia, 48*(4), 921–932. https://doi.org/10.1016/j.neuropsychologia.2009.11.011.

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). Springer. https://doi.org/10.1007/978-1-4419-6546-2_14.

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

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490–517. https://doi.org/10.1037/0033-295X.103.3.490.

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

Lehmann, M., Hagen, J., & Ettinger, U. (2022). Unity and diversity of metacognition. *Journal of Experimental Psychology: General, 151*(10), 2396–2417. https://doi.org/10.1037/xge0001197.

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General, 149*(9), 1788–1799. https://doi.org/10.1037/xge0000746.

Mazerolle, M. J. (2023). *Package "AICcmodavg."* https://cran.r-project.org/web/packages/AICcmodavg/AICcmodavg.pdf.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience, 33*(5), 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press.

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

Peng, Y., & Tullis, J. G. (2021). Dividing attention impairs metacognitive control more than monitoring. *Psychonomic Bulletin & Review, 28*(6), 2064–2074. https://doi.org/10.3758/s13423-021-01950-9.

Qiu, L., Su, J., Ni, Y., Bai, Y., Zhang, X., Li, X., & Wan, X. (2018). The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biology, 16*(4), e2004037. https://doi.org/10.1371/journal.pbio.2004037.

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*(7), e98663. https://doi.org/10.1371/journal.pone.0098663.

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001.

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. https://doi.org/10.1016/j.lindif.2012.12.003.

Roebers, C. M., Mayer, B., Bayard, N., Steiner, M., & 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.

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*(4), 749–767. https://doi.org/10.1348/978185409x429842.

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*(1), 110–128. https://doi.org/10.1080/15248372.2016.1157079.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience, 1*, e17. https://doi.org/10.1017/pen.2018.16.

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

Schneider, W., Tibken, C., & Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications. *Advances in Child Development and Behavior, 63*, 273–307. https://doi.org/10.1016/bs.acdb.2022.04.006.

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite chance task performance. *Psychological Science, 25*(12), 2199–2208. https://doi.org/10.1177/0956797614553944.

Selmeczy, D., & Ghetti, S. (2019). Here is a hint! How children integrate reliable recommendations in their memory decisions. *Journal of Experimental Child Psychology, 177*, 222–239. https://doi.org/10.1016/j.jecp.2018.08.004.

Selmeczy, D., Ghetti, S., Zheng, L. R., Porter, T., & Trzesniewski, K. (2021). Help me understand: Adaptive information-seeking predicts academic achievement in school-aged children. *Cognitive Development, 59*, 101062. https://doi.org/10.1016/j.cogdev.2021.101062.

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*(1), 3–27. https://doi.org/10.1007/s11409-019-09208-5.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290–302. https://doi.org/10.1037/0022-0663.86.2.290.

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

van Loon, M., De Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology, 158*, 77–94. https://doi.org/10.1016/j.jecp.2017.01.008.

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*(2), 173–191. https://doi.org/10.1007/s11409-013-9100-0.

van Loon, M. H., & Oeri, N. S. (2023). Examining on-task regulation in school children: Interrelations between monitoring, regulation, and task performance. *Journal of Educational Psychology, 115*(3), 446–459. https://doi.org/10.1037/edu0000781.

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.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences, 15*, 159–176. https://doi.org/10.1016/j.lindif.2004.12.001.

Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills: Metacognition in the numerical domain. *Psychological Science, 25*(9), 1712–1721. https://doi.org/10.1177/0956797614538458.

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

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition, 22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004.