Cognitive Development 74 (2025) 101561

Contents lists available at ScienceDirect

Cognitive Development

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

Metacognitive monitoring in early elementary school-aged children: Task dependency in monitoring judgments, task consistency in monitoring behaviours

Janina Eberhart [a,b,*,1], Kou Murayama [b,2], Michiko Sakaki [b,3], Donna Bryce [a,c,4]

[a] Department of Psychology, University of Tübingen, Schleichstraße 4, Tübingen 72076, Germany
[b] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, Tübingen 72072, Germany
[c] Department of Psychology, University of Augsburg, Universitätsstraße 10, Augsburg 86159, Germany

ARTICLE INFO

Keywords:
Metacognitive monitoring
Monitoring accuracy
Monitoring behaviour
Individual differences
Elementary school children

ABSTRACT

Children's metacognitive monitoring is typically considered as a domain general skill that can be applied in different tasks and situations. However, this assumption lacks empirical evidence as few studies tested whether children's accuracy of monitoring judgments as well as their monitoring behaviours are consistent across tasks. It is also not clear if children who provide more accurate monitoring judgments also show more frequent monitoring behaviours. In the current research study 53 elementary school children's metacognitive monitoring was assessed with four tasks: on the one hand, the accuracy of children's monitoring judgments was assessed with two computer-based tasks (one task required monitoring of memory and the other task required monitoring of reaction times); on the other hand, the frequency with which they engaged in monitoring behaviours was assessed with two construction tasks. Correlational analysis showed that there was no significant association between children's monitoring judgment accuracies. In turn, children's monitoring behaviour on two construction tasks was significantly positively associated. Intercorrelations between children's monitoring judgment accuracies and monitoring behaviours showed that children who more accurately monitored their reaction time showed significantly more monitoring behaviour when working on construction tasks. Conversely, children's monitoring judgment accuracy on a memory task was not significantly associated with their monitoring behaviour. These findings suggest that the processes underlying children's monitoring judgments may be task specific, whereas their tendency to engage in monitoring behaviours may be domain general. Implications for promoting metacognitive monitoring are discussed.

* Corresponding author at: Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany.
E-mail address: janina.eberhart@uni-tuebingen.de (J. Eberhart).
[1] ORCID: 0000-0002-3650-5452
[2] ORCID: 0000-0003-2902-9600
[3] ORCID: 0000-0003-1993-5765
[4] ORCID: 0000-0001-8311-4457

https://doi.org/10.1016/j.cogdev.2025.101561
Received 6 September 2024; Received in revised form 20 December 2024; Accepted 11 February 2025
Available online 27 February 2025

1. Introduction

On a daily basis children experience situations in which they want to master specific tasks. Especially when working alone, accurately evaluating one's own progress and performance in a challenging task can be beneficial as it allows one to adapt and select better strategies in order to succeed. For example, when completing homework, a child may ask themselves at the end of a paragraph whether they understood the text and they may also make judgments about their confidence in correctly answering specific questions. Both of these processes can be referred to as metacognitive monitoring, a component of the broader construct metacognition. Metacognitive monitoring has been operationalised in different ways, i.e., engaging in monitoring behaviours and the monitoring judgments participants provide. At this point it is unclear if metacognitive monitoring is a unitary construct in children which enables them to both engage effectively in monitoring behaviours and provide accurate monitoring judgments across different contexts. The current study investigates these different types of monitoring and their associations in a sample of elementary school children with the aim of uncovering the nature of metacognitive monitoring in young children.

1.1. Metacognition and the assessment and development of metacognitive monitoring in children

Metacognition can be broadly described as thinking about one's own thinking (Flavell, 1979). Nelson and Narens (1990) introduced a model of procedural metamemory processes, now more widely applied outside of the domain of memory, which differentiates between monitoring and control. Monitoring is a process that serves to update a person's representation of the task (the meta-level) such as reflecting on one's own understanding of a text; control includes actions taken to adapt situations or learning processes based on the contents of the meta-level such as re-reading the text (Nelson & Narens, 1990). As is implied by the above definition, monitoring is a cognitive process which cannot be directly observed and as such is operationalised and assessed in different ways.

Metacognitive monitoring has been assessed using think-aloud, self-report questionnaires, and prospective and retrospective monitoring judgments. However, for young children methods that rely heavily on language and require a fairly sophisticated meta vocabulary are less suitable (Whitebread et al., 2009), meaning monitoring judgments are typically used. In these tasks children complete a learning task such as a memory or general knowledge task and provide a judgment on aspects of their performance, often with the help of a pictorial scale (e.g., Destan & Roebers, 2015; Koriat & Ackerman, 2010). Previous studies show that children as young as 6-years can provide fairly accurate monitoring judgments about their own performance (Destan & Roebers, 2015). These tasks are highly standardized making them comparable across children; however, the method of providing judgments about their own performance is not often encountered by children in their regular environment and therefore lacks ecological validity. Further, it is not known whether children naturally make such monitoring judgments during a learning task or not. Another age appropriate way to assess children's metacognitive monitoring is by observing children's behaviour while they undertake a learning or problem-solving task (e.g., Bryce & Whitebread, 2012; Marulis & Nelson, 2021). For this approach specific behaviours such as checking a plan during a task are proxies for metacognitive monitoring and it can be argued that this approach better captures naturally occurring monitoring behaviour as children are not prompted to monitor.

Both measurement approaches, namely collecting monitoring judgments and observing monitoring behaviours, provide important insights into the development of metacognitive monitoring. In fact, child observations provided evidence that children's metacognitive monitoring abilities begin to develop earlier than it was long assumed (Whitebread et al., 2009). Whitebread et al. (2009) found that 3- to 5-year-old children showed monitoring behaviours (e.g., reviewing progress on task) in particular during child-initiated learning activities. A cross-sectional study that adopted observational coding with children aged 5 and 7 years found evidence for both quantitative increases in the frequency of monitoring behaviours as well as qualitative changes in the types of monitoring behaviours demonstrated with age (Bryce & Whitebread, 2012). Moreover, as of the age of four, children were able to indicate their confidence about recognizing previously learned picture pairs on a pictorial rating scale (Kälin & Roebers, 2020). In the elementary school years, 8- and 10-year-old children tend to overestimate themselves, but overall, they can make accurate judgments on their knowledge about specific learning material (Metcalfe & Finn, 2013). Koriat and Ackerman (2010) also found that 8-, 9- and 11-year-old children were able to accurately judge if their answers were correct or incorrect. Furthermore, they found that relative monitoring accuracy for the 11-year-olds was significantly higher than for the two younger age groups, indicating an age-related increase. While these studies clearly demonstrate that monitoring abilities begin to emerge in childhood, there is also evidence that they continue to mature into adulthood. Interestingly, in terms of monitoring judgment accuracy there is evidence that the developmental trajectory is moderated by the type of monitoring judgment. That is, von der Linden et al. (2016) observed that adults' confidence judgments were more accurate than 8-year-olds' confidence judgments, whereas the accuracy of judgments of learning was rather stable across four age groups (children, adolescents, younger adults, older adults).

Taken together, children's metacognitive monitoring can be reliably assessed in elementary school children with different measurement approaches (monitoring judgments and child observations) and there is evidence that monitoring processes become more sophisticated and accurate with age. Understanding the nature of metacognitive monitoring at the beginning of formal schooling is particularly important, as this is when formal instruction about metacognitive strategies or behaviours becomes possible, and it precedes the age range for which most school-based metacognition interventions are designed and implemented (Eberhart, Ingendahl, & Bryce, 2024). Indeed, Janssen and Lazonder (2024) observed that interventions to improve monitoring accuracy were descriptively more effective for elementary school students than for secondary school students. Therefore, the current study focused on this age group.

In this section different measurement approaches of metacognitive monitoring and how these were used to understand the development of metacognitive monitoring were described. Next, we examine studies that assessed metacognitive monitoring with the

same measurement approach, either monitoring judgments or child observation, in more than one task.

1.2. Stability of monitoring judgments and monitoring behaviours

Empirical studies suggest that the accuracy of metacognitive monitoring judgments may vary across judgment types and tasks (Mengelkamp & Bannert, 2010), although the majority of these studies were conducted with adults. The accuracy of monitoring judgments can be calculated in various ways and always takes objective performance into account. For example, the absolute accuracy is the difference between performance and judgments, and the relative accuracy is the correlation between performance and judgment (Mengelkamp & Bannert, 2010). It needs to be acknowledged that results can differ depending on how monitoring accuracy is calculated – for instance, absolute accuracy can be high and relative monitoring can be low or the other way around (Dentakos, Saoud, Ackerman, & Toplak, 2019). Importantly, by calculating the relative monitoring accuracy in the current study we focus on how sensitive children are to trial-by-trial variations in their performance because we reasoned this may be particularly relevant information that is used to drive subsequent control adjustments to behaviour. In a study with undergraduate students, Mengelkamp and Bannert (2010) examined the relative accuracy of monitoring judgments given at different time points in relation to a learning activity. They found that monitoring accuracy was not stable across judgments given before, during, and after a task. McDonough et al.'s (2021) study confirmed these findings and suggested that monitoring ability may vary depending on whether monitoring judgments are given at encoding or retrieval. In contrast, a small scale study by Desoete (2008) assessed third graders' prospective and retrospective monitoring judgments in relation to their performance on a math problem-solving task and they found significant correlations between the accuracy of these two types of judgments. However, these studies compared metacognitive monitoring judgments at different times (e.g., before and after responding to task items) but within the same task. The few studies that assessed monitoring accuracy with different tasks found mixed results. Mengelkamp and Bannert (2010) observed that the accuracy of students' confidence judgments on two separate tests related to either text comprehension or transfer of knowledge to a new scenario were significantly associated. In contrast, Kelemen, Frost, and Weaver (2000) examined different types of relative monitoring accuracy in four memory tasks in a sample of undergraduate students. The memory tasks concerned students' ability to remember Swahili-English word pairs, unrelated word pairs of English nouns, answers to general knowledge questions, and recalling material from a text. While they found that students showed similar levels of memory recall and confidence across tasks, monitoring accuracy was not correlated. Similarly, Dentakos et al. (2019) assessed undergraduate students' relative monitoring accuracy in domains such as general knowledge, financial calculation, probability calculation, and emotion recognition. In each domain participants answered multiple choice questions and provided retrospective confidence judgments of each individual item. Intercorrelations of the relative monitoring accuracy scores across domains indicated no significant associations, although participants did show a consistent tendency to over- or under-estimate themselves across domains (Dentakos et al., 2019).

While these studies are certainly informative, the findings from adults cannot be directly translated to children. Given that the accuracy of monitoring judgments remains an oft-used measure of metacognitive monitoring in developmental studies (Koriat & Ackerman, 2010; Metcalfe & Finn, 2013), it seems pertinent to also investigate this issue directly in young learners. Interestingly, to date few studies have examined differences in how children and adults form monitoring judgments outside the realm of memory tasks and judgments of knowing (see Schneider & Lockl, 2008 for a review of this literature). Those that do exist, however, provide evidence that confidence judgments made by children and adults differ in important ways. For instance, children's judgments are more affected by the type of question they are asked, with their monitoring abilities suffering when faced with misleading or biased questioning whereas adults' confidence judgments are rather robust in comparison (Roebers, 2002; Roebers & Howie, 2003). Further, 9-year-old children make better use of information from retrieval processes (similar to adults) when forming confidence judgments than do 7-year-olds (Roderer & Roebers, 2010); this implies that throughout childhood monitoring judgments are formed more consciously and more valid sources of information contribute to judgments. In an intervention study designed to enhance the use of recollection-based cues, von der Linden et al. (2016) observed that the accuracy of retrospective confidence judgments made by participants from four age groups (children, adolescents, younger adults, older adults) benefitted from the intervention, whereas prospective judgments of learning only became more accurate after the intervention for the child group. The authors speculate these developmental trends may be due to the different types of monitoring judgments drawing on different sources. As such, it is feasible that young learners' monitoring judgment accuracy could be stable across tasks and judgment types, even though adults' appear not to be, if they for example rely on different cues and sources when forming their monitoring judgments. The current study aims to address the research gap regarding studies with children to understand children's monitoring judgments across tasks and judgment types.

Even fewer studies have assessed naturally occurring monitoring behaviour in more than one task. Therefore, it remains an open question if children who engage in frequent, unprompted monitoring behaviour in one task also do so in another task. One exception is a study by Spektor-Levy, Basilio, Zachariou, and Whitebread (2017) who assessed elementary school children's metacognitive monitoring with behavioural coding in two different construction tasks. In one construction task, children were asked to build different shapes with train tracks, and in the other construction task, children were asked to build different models with LEGO bricks. For both tasks, the shape / model where it was judged that children had faced an appropriate challenge was selected for behavioural coding. That is, the number of times children demonstrated monitoring behaviours was counted. In that study, while children who monitored themselves more frequently performed better on the tasks, there were no associations between frequency of monitoring behaviour across the two tasks. Clearly, further evidence is needed before strong conclusions can be drawn about whether engaging in unprompted monitoring behaviours is rather consistent within an individual or very dependent on the task.

To sum up, currently it is unknown whether there is consistency within one child and whether children who make highly accurate monitoring judgments in one task also make highly accurate monitoring judgments in another task. Similarly, it is not clear if children

who frequently engage in monitoring behaviours in one task also do so on another task. Next, we will highlight studies that included multiple indicators of monitoring.

1.3. Associations between different indicators of monitoring

Most studies that aim to assess monitoring only use one indicator of monitoring, and the choice of indicator is often influenced by the research tradition they are situated in. That is, researchers from a cognitive psychology background tend to opt for prompted monitoring judgments, and researchers from an educational psychology background tend to focus on unprompted, monitoring behaviours and employ self- / teacher-report or observational methods (Kim, Zepeda, & Butler, 2023). Self- / teacher-reports typically ask about general monitoring behaviours, whereas observational methods are more task-specific. Given this, few studies have incorporated multiple indicators of monitoring, and there is a particular scarcity of such studies conducted with children. One exception is the previously mentioned study with third graders by Desoete (2008) who applied a variety of measurement approaches including teacher-reports, self-report questionnaires, think-aloud, and prompted monitoring judgments. Focusing on their results related to monitoring, the study indicated that children who were rated by teachers to engage more often in monitoring also evidenced more metacognition in the think-aloud protocols. Händel and Dresel (2022) asked undergraduate students to report on the frequency of their monitoring strategy use on the one hand, and to provide item-specific monitoring judgments during a knowledge test on the other hand. Interestingly, they found no or only weak significant associations between students' self-reported frequency of monitoring strategy use and the accuracy of their monitoring judgments. This is surprising as a logical argument would be that engaging more often in monitoring should provide more awareness of one's own performance. In turn, an increased awareness of one's performance together with the experience of feedback on one's action (e.g., a grade in educational contexts), should help to calibrate the accuracy of one's monitoring, resulting in more accurate monitoring judgments. Given recent calls to better integrate across research traditions (Kim et al., 2023), it seems pertinent to evaluate the commonalities among different indicators of metacognitive monitoring, as it is currently unknown to what extent findings based on one indicator can be generalised to another.

1.4. The current study

There is a paucity of empirical evidence regarding whether children show consistent monitoring judgments and monitoring behaviours across tasks. Further, it is not known if children who provide more accurate monitoring judgments when prompted also show more frequent spontaneous monitoring behaviour. The present pre-registered study[5] addresses these two research questions concerning children's metacognitive monitoring across four tasks (two tasks with monitoring judgments and two tasks with monitoring behaviour). Children's metacognitive monitoring skills were assessed directly (i.e., as they pertain to a specific task) using methods that are commonly used for measuring metacognitive monitoring in children. Furthermore, on all four tasks children's metacognitive monitoring was measured during task processing and locally – with trial-by-trial judgments for the tasks with monitoring judgments and observational coding during the task for the monitoring behaviours. As such we gain insights into children's ability to monitor themselves in-the-moment rather than how they make global metacognitive judgments which may reflect different processes (Händel, de Bruin, & Dresel, 2020). None of the tasks required any of the classical academic skills such as reading, writing, or math. To shed light on our proposed research questions, we selected two pairs of tasks that were similar in terms of their general structure but varied in terms of the primary task. In two highly structured, computer-based tasks, children were prompted to provide monitoring judgments after each trial and the relative accuracy of these judgments was calculated. These tasks were chosen as they had a similar structure, were novel to children, and it was unlikely that children had prior experience monitoring themselves on such tasks. However, the tasks also differed as one task was a working memory task and the other task was an inhibitory control task. Furthermore, in the working memory task children were asked to provide a monitoring judgment on the accuracy of their recall whereas in the inhibitory control task children had to judge the reaction time of their response. Thus, both primary tasks tapped an aspect of children's higher-order cognitive skills, typically referred to as executive functions. Children's spontaneous monitoring behaviours were assessed using observational coding of their behaviour while they completed two construction tasks. Common to both of these tasks is the fact that children may have had experience with the task materials, but the task demands were novel. The tasks have been used previously in the field with similar age groups and were chosen as they are challenging for children of this age group which ensures variation in performance and the need to monitor. The tasks also differed to a certain extent, for example one task is only built in two dimensions and using a black and white symbolic plan whereas the other is built in three dimensions using a more detailed coloured plan.

The research questions addressed in the present study are described here and an illustration can be found in Fig. 1.

(1) Is the accuracy of children's monitoring judgments consistent across two tasks? Is the frequency of children's monitoring behaviours consistent across two tasks? We investigated whether children who give more accurate prompted monitoring judgments in one computer-based task also give more accurate prompted monitoring judgments in the other computer-based task (see Fig. 1, RQ 1a). Studies with adults that examined associations between the accuracy of monitoring judgments on different tasks reported mixed results (Dentakos et al., 2019 found no associations; Mengelkamp & Bannert, 2010 found associations). It was also evaluated whether children who more often spontaneously monitor on one construction task also more often

[5] This study was preregistered on OSF, please refer to this OSF link: https://doi.org/10.17605/OSF.IO/N8ZQH

spontaneously monitor on the other construction task (see Fig. 1, RQ 1b). The only study that assessed monitoring behaviours on two tasks found no significant associations (Spektor-Levy et al., 2017). Thus, given the lack of a pattern in previous studies, no hypotheses were formulated.

(2) Is the accuracy of children's monitoring judgments associated with the frequency of their monitoring behaviours (see Fig. 1, Research Question 2)? No study to our knowledge explored the association between accuracy of monitoring judgments and frequency of monitoring behaviour. A study that assessed undergraduate's accuracy of monitoring judgments on the one hand and their self-reported frequency of monitoring strategy use on the other hand, indicated only low or no significant associations (Händel & Dresel, 2022). Little is known about the association between children's accuracy of monitoring judgments and the frequency of their monitoring behaviours. Thus, no hypothesis was formulated.

2. Methods

2.1. Participants

Sixty-two children were recruited to participate in the study. Eight children did not want to participate on the day of data collection, and one child was excluded as their chronological age did not match their school year and as such they were outside our targeted age group. Thus, the sample of the current study included 53 children ($M_{age} = 7.38$ years; $min_{age} = 6.29$ years; $max_{age} = 8.67$ years; 50 % female) who attended first and second grade in four elementary schools. All schools were located in a small University town in Germany. Parents reported on the languages spoken at home and the highest educational level for both parents. Parent reports indicated that for 85 % of the children German was the first language and 33 % of the children spoke a second language at home. Most children came from highly educated families. Fifty out of 53 families provided information about the highest educational level of parent 1 and it was indicated that 74 % had a College or University degree; 46 out of 53 families provided information about the highest educational level of parent 2 and it was indicated that 85 % had a College or University degree. This study was approved by the ethical review board of the Faculty of Economics and Social Sciences at the University of Tübingen.

Two approaches were taken to estimate the required sample size for this study. Previous studies that assessed metacognitive skills in elementary school children with at least two approaches revealed correlations ranging from $r = .21$ to $r = .55$ (mean $r = .38$). Based on this information, we first completed an a priori power analysis with an $\alpha = .05$, $\beta = .80$ (two-tailed test) and a range of true correlations from $r = .15$ to $r = .50$ (the required sample sizes ranged from $N = 343$ to $N = 26$). Secondly, we conducted a more specific data simulation in which data was generated based on the mean (and standard deviation) monitoring accuracy and monitoring frequency observed in our own previous studies. For these data simulations, true correlation coefficients ranging from .30 to .40 were used and power was calculated as the proportion of studies that gave a true positive result. To have power to observe a correlation of $r = .38$ (the mean value of previous studies), we aimed for a sample size of 55 children. Please refer to our preregistration for more details on our power analyses (seehttps://doi.org/10.17605/OSF.IO/N8ZQH).

2.2. Measures

2.2.1. Metacognitive Odd One Out task

The Odd One Out (adapted from Alloway, 2007) is a computer-based working memory task which has been used widely with 6- to 7-year-olds (de Abreu, Puglisi, Cruz-Santos, Befi-Lopes, & Martin, 2014; Henry, Messer, & Nash, 2014). In this metacognitive version of the task, children were shown three boxes each containing a shape: two shapes are the same and one is different (Odd One; see Fig. 2A) and children had to identify the odd shape by clicking on it with the computer mouse. They were also instructed to memorize the location in which the odd shapes were presented. After making a sequence of five[6] Odd One judgments, they were asked to recall the location of the Odd Ones in the correct order. After each trial, children were prompted to provide a confidence judgment where they indicated on a continuous visual analogue scale (VAS) labelled with smileys (see Fig. 2A) how confident they were that they had recalled the order and locations of the Odd Ones correctly.

The task was programmed in PsychoPy (Peirce et al., 2019) and completed on a laptop (Lenovo, T450s and T470s, 13 in. screen) with a mouse attached. For each screen in which an odd shape was to be identified two out of a possible 12 geometric shapes were randomly selected, as was which of the three boxes contained the odd shape. Throughout the task, when the child clicked on a box with their mouse it turned grey for 500 ms. When they clicked on the VAS to provide their confidence judgment, a red circle appeared on the scale. To progress through the experiment, children always clicked on a green rectangle on the lower part of the screen.

At the very start of the task, a screen was presented with images of well-known characters from children's books and television. This allowed the research assistant to observe how proficient the child was using the mouse, as when the child clicked on the character it disappeared. Next followed a series of practice tasks. First, children completed three practice trials where they identified and recalled the odd shapes. After each trial children received feedback on their performance. Second, to introduce children to the confidence judgments and the use of the VAS, we applied an approach that has been developed by Roebers, Cimeli, Röthlisberger, and Neuenschwander (2012). Children answered three questions which varied in their difficulty level (e.g., "how old are you?" vs. "how much hair do you have on your head?"). After each question, children were asked to provide a confidence judgment ("how certain are you

[6] Sequence length was selected based on pilot testing with children in the same age range, aiming to achieve approximately 70 % proportion of the sequence correctly recalled.

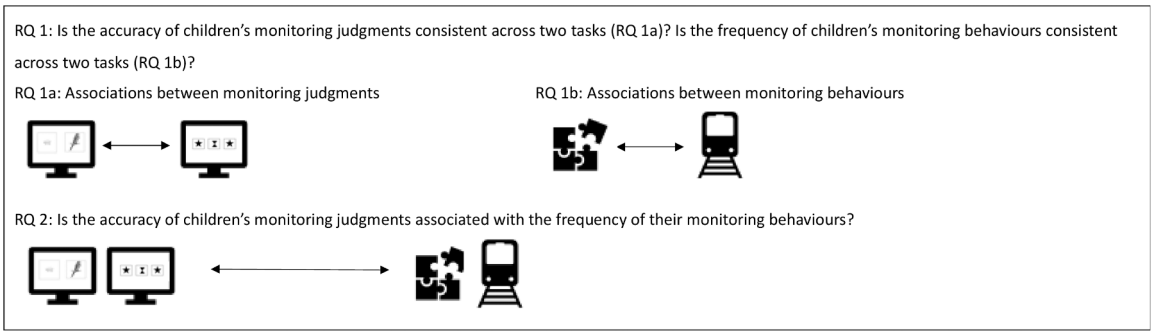

Fig. 1. Visualization of Research Questions 1 and 2.

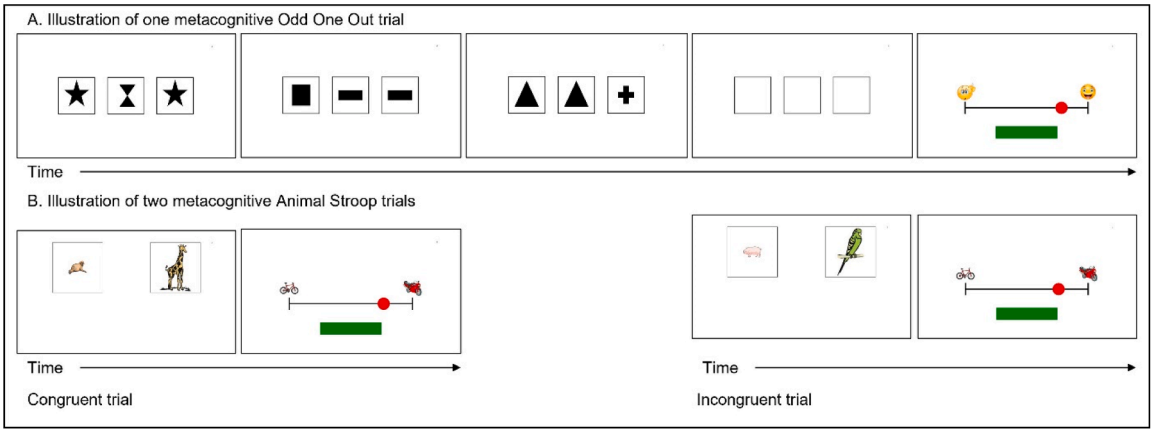

Fig. 2. Illustration of the Odd One Out Task (A) and the Animal Stroop Task (B).

about your answer?"). Finally, children completed three practice trials of varying difficulty (sequence lengths) including the identification and recall of the odd shape as well as the confidence judgments. Once the practice trials were completed, and the research assistant was sure that the child knew how to use the VAS, the experimental trials started. The child initiated the start of each trial by mouse-click. Children completed ten trials with a sequence length of five and two trials of a sequence length of two. The trials with a sequence length of two were added for motivational purposes and were not analysed. On each trial, children's percentage of correct recall could thus be 0 %, 20 %, 40 %, 60 %, 80 % or 100 %. After every trial, children provided a confidence judgment on a continuous VAS and the location of their click was converted to a numerical value ranging between 0 (very unsure) and 1 (very sure). The dependent variable (calculated for each child) is relative monitoring accuracy, that is, the Pearson product-moment correlation between their confidence judgments and the percentage of correct recall (Mengelkamp & Bannert, 2010). A stronger positive correlation indicates children are more sensitive to trial-by-trial variations in their performance. If a child had zero variance on one of the two variables (confidence judgments or percentage of correct recalls), the child had to be excluded from this analysis.

2.2.2. Metacognitive Animal sSroop task

In the Animal Stroop task (Bryce, Szűcs, Soltész, & Whitebread, 2011) children are presented with two pictures of animals on a screen next to each another, one of which is presented physically larger on the screen. Children should click as fast as they can on the animal that would be larger in real life. In congruent trials the animal that is larger on the screen is also larger in real life. In incongruent trials the animal that is larger on the screen is smaller in real life (see Fig. 2B). In this metacognitive version of the Animal Stroop, at the end of each trial children were asked to estimate on a VAS labelled with a bicycle and a motorcycle how quickly they had responded (estimated Reaction Time = eRT; see Fig. 2B).

The task was also programmed in PsychoPy (Peirce et al., 2019) and completed using the same equipment as mentioned above. In total there were 36 animal images used in this task, 12 that are small-sized, 12 that are medium-sized and 12 that are large-sized in real-life. On each trial, two animals of different real-life sizes were selected randomly from the relevant animal images and presented on the screen within black frames around them. One animal image was presented smaller on the screen (15 % of the height of the laptop screen) and the other animal image was presented larger on the screen (40 % of the height of the laptop screen). The task included easy congruent (i.e., the animal that is larger on the screen is also larger in real life than the other animal on the screen) and difficult incongruent trials (i.e., the animal that is larger on the screen is smaller in real life than the other animal on the screen). The varying difficulty of congruent and incongruent trials created variability in reaction times which ensured a more reliable assessment of

monitoring accuracy. When they clicked on the VAS to provide their estimate of their reaction time, a red circle appeared on the scale. To confirm their estimate and to initiate the start of a new trial, children clicked on a green rectangle on the lower part of the screen.

Similar as in the Odd One Out task, at the very beginning of the task children were presented some images with animals to click on to practice the use of the computer mouse. Next, children were introduced to the task with two practice trials and received an explanation that they should click as fast as possible on the animal that would be larger in real life. The two practice trials consisted of a congruent and incongruent trial. Children received feedback after each trial. Next, the VAS was explained to the children. To practice its use children were shown three labyrinths of varying complexity on the screen. Children were asked to move their mouse from a designated starting point to the end point through the labyrinth. Three labyrinths were included that varied in the number of blockages encountered, and therefore the time taken to complete it. After each labyrinth, children provided a judgment on the VAS about how long they had taken to reach the end of the labyrinth, their eRT. Then, children completed four trials of the Animal Stroop and provided estimates of their own reaction time afterwards. Each child completed 24 trials, half congruent and half incongruent, presented in a random order. After every trial, children provided an estimate of their RT on the continuous VAS labelled with a bicycle and a motorcycle and the position the child clicked was converted to a value ranging from 0 to 1.

For analysis, children's estimates were reversed, so that a higher value reflected longer estimated reaction times (eRTs). It should be noted that in a deviation from our pre-registered analyses, trials in which the child responded incorrectly (6 % of all trials) were excluded from further analysis due to the observation that children often confounded their eRT with the accuracy of their answer; it can be assumed that different metacognitive processes are involved in monitoring errors. The dependant variable (calculated for each child) was again relative monitoring accuracy, calculated as the Pearson product-moment correlation between estimated and objective RTs (Mengelkamp & Bannert, 2010).

2.2.3. Train Track task

This quasi-naturalistic, construction task has been used widely in the field (e.g., Bryce & Whitebread, 2012; Spektor-Levy et al., 2017). In this implementation of the task, children built two constructions one after the other using wooden train track pieces. Children were videoed during the task and their behaviour was coded at a later timepoint.

The behaviours of the children completing the task were recorded using video cameras and a tripod. Task materials included a range of wooden train track pieces and a train (see Fig. 3A) and laminated images of the shapes to be built (see Fig. 3B). These shapes were selected based on a pilot study conducted with five children in 1st and 2nd grade. Two shapes of a reasonable challenge which were neither too easy nor too difficult for children of this age were selected. During data collection, the research assistants also had an easier and a harder shape available to them in case the selected shapes were too easy (no monitoring required) or too hard (signs of frustration rather than metacognition). These were, however, not required in this study.

At the start of the task the wooden train track pieces were in a pile on the floor next to the child (i.e., not organised as in Fig. 3A). This procedural decision allows researchers to observe certain monitoring behaviours (e.g., reviewing the different pieces) that might be missed if materials are pre-sorted for the children. The task was introduced to the children by presenting the task material (train track pieces and laminated images of the train track shapes) to the children. Then, children were told that they will be asked to use the train track pieces to build a shape that looked exactly as the shape on the picture. Children built two shapes. They started with the "Goggles"-shape and continued with the P-shape (see Fig. 3B). These names were not used when the shapes were introduced to the child. No practice trials were included. Children completed the task without support from the researcher and determined themselves when they had completed the task (up to a maximum of 4 min). If the child sought help from the research assistant, they responded with agreed-upon phrases, such as "Hm…it can be tricky. Let me know when you are done". If the child had not ended the task themselves within 4 min, research assistants offered the child to stop this attempt and move onto the next one. For 2 out of 108 constructions children took longer than 5 min to complete the task (maximum 5 min 43 s). Children were not informed of this time limit, in order to avoid creating the impression that speed was important for task success.

Observational video coding was applied to evaluate children's spontaneous, unprompted monitoring behaviour, following the procedure and using all monitoring codes from the scheme described in Bryce and Whitebread (2012). Please refer to Table 1 for descriptions and examples of all metacognitive monitoring codes, as well as the rationale for why each behaviour reflects metacognitive monitoring.

It is important to note that in the original version of this coding scheme, a separate category of codes describes behaviours thought to reflect metacognitive control. As such, with our coding we aim to only capture behaviours that reflect the child updating their mental representation of the task, and not the subsequent control of behaviour. Two undergraduate psychology students employed as research assistants participated in an intensive coder training and completed the coding. The coder training involved a comprehensive introduction to the coding scheme and the rationale for interpreting each behaviour as an indicator of monitoring (please refer to Table 1 for an overview of all codes) where each code was demonstrated with a sample video. The coders then coded four videos from this study independently using the coding scheme with the examples and explanations from Table 1 as reference. Subsequently, their codes were compared, and disagreements were discussed coming to an agreement on how to code behaviours. Coders were instructed to first watch the entire video and then to slowly work their way through the video implementing the coding. Coders had to make two coding decisions: first, they had to decide and annotate when in the video stream children showed monitoring behaviour; second, they had to decide which monitoring code (see Table 1) to assign to this behaviour. The inter-rater reliability between two coders was calculated based on double coding of a random sample of 10 % of videos (i.e., 20 videos). Two interrater reliability scores were calculated: Unitizing agreement which reflects whether the coders agreed that a certain section of the video stream showed monitoring at all (i.e., the first coding decision described above; 59 % agreement) and Coding agreement which reflects whether coders agreed on the code assigned to a certain behaviour (i.e., the second coding decision described above; Cohen's Kappa $k = .94$). After acceptable

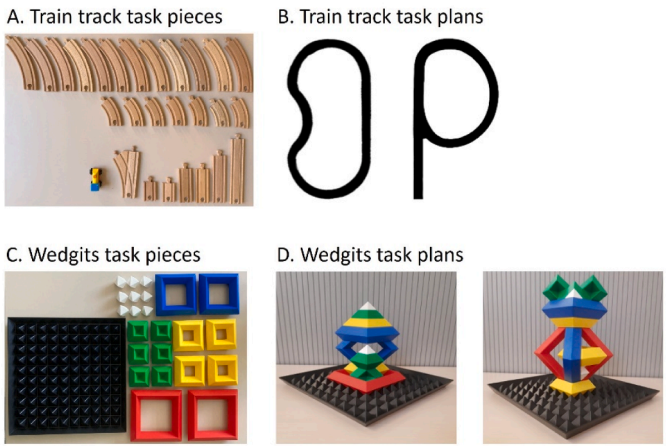

Fig. 3. The materials used in the Train Track Task (A, B) and in the Wedgits Task (C, D).
Note. The train track shapes in panel B are referred to in the manuscript as Goggles and P-Shape, and the Wedgits shapes in panel D are referred to in the manuscript as Mushrooms and Frog. These names were not communicated to the children. The Wedgits plans were presented to children in colour.

Table 1
Overview of the monitoring codes adapted from Bryce & Whitebread (2012).

Code name and description	Example	Rationale
Checking Own A pause to review whole of own construction. Not checking only one area. Can be accompanied by verbalisation.	Child pauses and looks at whole construction they have made so far. Not just looking at one part of the construction.	Child is updating their mental model of the task. The described behaviour may be an attempt by the child to answer the question "what have I built so far?"
Checking Plan Checking the original plan. Can be accompanied by verbalisation.	Child looks at the plan of the construction they are working on.	Child is updating their mental model of the task. The described behaviour may be an attempt by the child to answer the question "what needs to be built?"
Prospective Monitoring Judgment of the task before child begins working on the task. Always verbal.	*This is going to be a challenge!*	The outcome of a monitoring process. Prospective monitoring is reflected by a verbal judgment of task difficulty and/or own abilities.
Clarification Clarifying task demands with adult. Can be before or during task. Always verbal.	*Do I use all the pieces?*	Child is updating their mental model of the task. The described behaviour may be an attempt by the child to answer the question "what are the rules?"
Reviewing A pause to look purposefully at different pieces before or during task, not seeking any particular piece. Reviewing must last for at least 3 s. Can be accompanied by verbalisation.	Child looks around all the Train Track pieces/ Wedgits puzzle pieces (for at least 3 s).	Child is updating their mental model of the task. The described behaviour may be an attempt by the child to answer the question "what materials are available?"
Self-questioning Child highlights a problem to be solved, poses themselves a question. Always verbal.	*How will it curve around?* (to self)	Can either reflect that a child initiates a monitoring strategy or the outcome of a monitoring process – child has noticed an error / difficulty.
Commentary Comment on success, what has been achieved so far. Always verbal.	*There, that's a tiny bit better.* *Right, that bit is done…*	The outcome of a monitoring process – child is tracking their progress.
Evaluation Spontaneous evaluation of the final product at the end of the task. Can be verbal or nonverbal.	*But that bit isn't right.* Child looks at the finished construction and smiles / furrows their brows and frowns.	The outcome of a monitoring process – child has evaluated their own performance.
Justified Termination Child announces task is finished without being prompted at an appropriate point (i.e., the construction resembles the plan). Always verbal.	*There.* *Finished!*	The outcome of a monitoring process – child has tracked progress and compared this to a goal.

Note. This is the coding scheme that coders used for coding children's metacognitive monitoring behaviour; sentences in italics are examples of verbal utterances that children may have used. The Rationale column provides the rationale for why this behaviour is considered to reflect metacognitive monitoring behaviours.

inter-rater reliability was established, each coder coded half of the videos. Children received a monitoring score based on spontaneous monitoring behaviours as assessed with the coding scheme. To account for the duration that participants engaged with a task, the frequency count of their monitoring behaviour was divided by the time they have spent on the task (following the procedure

established in Bryce & Whitebread, 2012). For analysis a composite rate, namely the mean monitoring rate across the two train track shapes, was entered into analysis.

2.2.4. Wedgits puzzle task

Like the Train Track Task, the Wedgits puzzle task is a quasi-naturalistic construction task that has previously been used to assess metacognition in young children (Marulis & Nelson, 2021). In this 3D building task children are given images of puzzles to build on top of a black tray. To record the behaviours of children completing the task, the same video cameras and tripods were used as already mentioned. The task materials can be seen in Fig. 3C, and the primary plans that children completed are in Fig. 3D. The same piloting procedure as for the Train Track Task was followed and a set of plans of puzzles of varying difficulty level were identified. Out of the piloted plans of puzzles, two shapes that provided a suitable challenge were chosen. The research assistants also had one easier and one harder Wedgits plan available to use during data collection; for one child the easier plan was used (a simpler tower structure with only one perpendicular piece).

Research assistants introduced the task by putting the task material (black tray and colourful pieces) next to the child on the floor (i.e., not organised as in Fig. 3C) and showing some laminated images with shapes to the child. Children were told that they would receive an image of a shape which they should build with the pieces. Similar to the Train Track Task, children completed two puzzles: a "Mushroom"- and "Frog"-shape (see Fig. 3D) and these names were not used in the interaction with the children. The same procedure as for the Train Track Task was followed and children completed the task independently and self-determined when they completed the task. As previously mentioned, research assistants encouraged children to finish the task after 4 min. However, ten children took longer than 5 min to complete the task (maximum 9 min and 21 s).

The same coding scheme was applied to the videos of children completing the Wedgits puzzle task as described for the Train Track Task, also by two trained coders. Thus, the same monitoring behaviours were coded in the Wedgits task (please refer to Table 1 for an overview of all monitoring codes). Their inter-rater reliability was calculated based on double coding of a random sample of 10 % of videos (a more detailed description of the coder training and establishment of the inter-rater reliability can be found in the section above on the Train Track Task). As for the Train Track Task, two interrater reliability scores were calculated. Unitizing agreement[7] was 73 % and Cohen's Kappa was $k = .99$. A rate of monitoring per minute for each task was calculated as described above, and composite rates (i.e., mean of the two rates) were used in further data analyses.

2.3. Enjoyment

After each task, children were asked to report how much they enjoyed the tasks to inform our interpretations regarding overall motivation for the tasks. Children were shown a scale with smiley faces, and they were asked "How much fun was this game for you?". With the help of the smiley faces children indicated if they had "a lot of fun" (score of 1), "fun" (score of 2), "not so much fun" (score of 3) or "no fun at all" (score of 4).

2.4. Procedure

Parental consent forms with information about the study were sent home to all families through children's class teachers. Only children whose parents provided written consent were invited to participate. Research assistants provided a description of the tasks in child-friendly language and children were asked for verbal assent based on this information prior to data collection. Most children completed all four tasks described above in one session with a duration of approximately 45 min. However, if participants expressed signs of fatigue or if it was not possible for organizational reasons, the tasks were completed in two separate sessions. The computer-based and construction tasks were completed in a counterbalanced order as two separate blocks in a quiet room of the children's schools. The order of the tasks within the blocks was also counterbalanced. All tasks were administered one-to-one in person by a trained research assistant. Most of the children participated in the afternoon after their regular school day.

2.5. Data analytic plan

This study's design, research questions, and analysis plan were preregistered; see https://doi.org/10.17605/OSF.IO/N8ZQH. All children completed at least two tasks. If children only completed two computer-based or two construction tasks, their data contributed to RQ 1. If participants had complete data for a computer-based and a construction task, their data contributed to RQ 2. A participants' data on one task was excluded if there was evidence in the data of non-compliance or failure to understand the task. In the Animal Stroop task children chose between two animal pictures. Thus, there was a 50 % chance to provide an accurate response on each trial. Therefore, non-compliance was defined as an overall accuracy level less than 60 %. In the Odd One Out task children had to identify the odd shapes out of three boxes. Thus, there was a 33 % probability to identify the odd shapes by chance. Therefore, non-compliance

[7] It is notable that the unitizing agreement is higher for the Wedgits Puzzle Task than for the Train Track Task. Based on the experience gained in this study, we believe this is primarily due to the need for children to more closely examine the complex plans in the former task, making codable behaviour easier to identify. This may reflect a limitation of the Train Track Task. However, even though the agreement differed across the two tasks, the monitoring rates for the two tasks were positively correlated, suggesting that coded behaviours in the two tasks still reflect a similar psychological construct.

was defined as correctly identifying the deviant shape in less than 40 % of all judgments in the experiment. For the Train Track and Wedgits tasks non-compliance would be reflected in children refusing to complete the task. In cases where a research assistant utilized one of the alternative shapes in the construction tasks, the child's behavior in the replaced task (deemed too easy or too hard) was not coded. Data were also excluded if there was evidence of a ceiling effect on task performance. In the Animal Stroop task this was not applicable as our primary interest was reaction times. In the Odd One Out task this would be if a participant recalled every single sequence correctly.

To address RQ1a, whether children who give more accurate prompted monitoring judgments in one computer-based task also give more accurate prompted monitoring judgments in the other computer-based task, the Pearson correlation between the relative monitoring accuracy scores was calculated. To address RQ1b, whether children who more frequently monitor their behaviour on one of the construction tasks do so as well on the other construction task, the Pearson correlation between the monitoring rate (instances of monitoring behaviour per minute) was calculated. The goal of RQ 2 was to assess if the children's accuracy of monitoring judgments and their monitoring behaviours were associated. For this, intercorrelations between participants' relative monitoring accuracy and their monitoring behaviour frequency were calculated.

3. Results

Descriptive statistics and correlations for all measures are presented in Table 2. Most children (89 %) had complete data on all four tasks. The data of three children had to be excluded from the Animal Stroop and Odd One Out analysis due to technical difficulties, lack of motivation, or difficulties understanding the task. Two children chose not to complete the Animal Stroop task, four children aborted the Odd One Out task before completion and one child chose not to complete the Wedgits puzzle task. One child only completed one of the two train track shapes, and another child only completed one of the two Wedgits shapes; in these cases, the monitoring rate and quality score from the completed shape was entered into analyses.

Monitoring accuracy (i.e., the correlation between children's objective performance and their monitoring judgments) was significantly different from zero for the Animal Stroop task ($N = 48$; $\bar{r} = .42$; $p < .001$) and the Odd One Out task ($N = 45$; $\bar{r} = .36$; $p < .001$; see Table 2). A paired-samples t-test indicated no significant differences for monitoring accuracy between these two computer-based tasks, $t(43) = -0.65$, $p = .520$, $d = .17$. In terms of the construction tasks, the monitoring rate was generally higher during the Wedgits puzzle task ($N = 52$; $M = 7.07$) than during the Train Track Task ($N = 53$; $M = 4.28$), which was confirmed by a paired-samples t-test, $t(51) = -9.77$, $p < .001$, $d = 1.26$. Overall, children showed similar levels of enjoyment on all tasks, $F(3, 192) = 0.39$, $p = .758$, $\eta^2 = .01$. Given the narrow age range, unsurprisingly, age was not significantly associated with any of the dependant measures.

3.1. Correlational analysis to assess research Question 1

In RQ 1a we asked whether children who give more accurate monitoring judgments when prompted in one computer-based task also give more accurate monitoring judgments when prompted in the other computer-based task. To address this research question, we calculated the Pearson correlation between the monitoring accuracy on the computer-based tasks, namely the Animal Stroop and Odd One Out tasks. The result of this correlation indicated a non-significant association ($N = 44$; $r = .01$; $p = .946$; see Fig. 4).

RQ 1b poses the question whether children who more often spontaneously monitor on one construction task also more often spontaneously monitor on the other construction task. A Pearson correlation between the monitoring rate on the Wedgits puzzle task and the Train Track Task indicated a significant, positive association ($N = 52$; $r = .57$; $p < .001$; see Fig. 4).

3.2. Correlational analysis to assess research question 2

To address RQ 2, we explored if children's monitoring accuracy on the computer-based tasks is positively related to the rate of their spontaneous monitoring behaviour during the construction tasks. As a significant positive association was observed between children's monitoring behaviour, a composite score was calculated as the mean of the monitoring rate of the Train Track and Wedgits puzzle tasks. Since no significant correlation was observed for children's monitoring accuracy, no composite score was created.

No significant association was observed between children's monitoring accuracy score on the Odd One Out task and children's monitoring rate on the construction tasks ($N = 42$, $r = .17$, $p = .274$). However, there was a significant, positive association between children's monitoring accuracy score on the Animal Stroop task and the monitoring rate on the construction tasks ($N = 45$, $r = .46$, $p = .001$; see Fig. 5).

4. Discussion

In the current study, the nature of metacognitive monitoring in young children was examined, with a particular focus on the task-specificity of monitoring judgments and monitoring behaviours, and the associations between these two indicators of monitoring. To this end, children aged between 6 and 8 years completed four tasks that exemplify typical approaches to assessing monitoring in young children. In two of these tasks, children were prompted to provide monitoring judgments and the accuracy of these judgments was calculated, in the other two tasks evidence of children's unprompted monitoring behaviours was observed and recorded. Correlational analyses indicated that while the accuracy of children's monitoring judgments was not consistent across the two tasks in this study, their tendency to engage in monitoring behaviours was. Further, the accuracy of one type of monitoring judgment (related to response

Table 2
Means, standard deviations, and correlations among all variables with confidence intervals.

Tasks and Variables	M	SD	1	2	3	4	5	6	7	8	9	10	11	12
Animal Stroop														
1. RT (s)	2.43	0.78												
2. eRT[1]	0.22	0.16	−.06 [−.34,.23]											
3. Monitoring Accuracy (Pearson's *r*)	0.42	0.29	−.34* [−.57, −.06]	.11 [−.18,.38]										
4. Enjoyment[2]	1.32	0.56	.24 [−.05,.49]	.10 [−.19,.38]	−.14 [−.41,.16]									
Odd One Out														
5. Proportion of correctly recalled odd shapes	0.57	0.16	−.29 [−.54,.01]	.22 [−.08,.48]	.21 [−.10,.47]	−.07 [−.37,.23]								
6. Mean CJ[3]	0.66	0.16	.13 [−.17,.41]	−.39** [−.62, −.11]	−.11 [−.39,.20]	.06 [−.24,.35]	−.07 [−.36,.22]							
7. Monitoring Accuracy (Pearson's *r*)	0.36	0.36	−.13 [−.41,.17]	.28 [−.02,.53]	.01 [−.29,.31]	−.05 [−.35,.25]	.33* [.05,.57]	.15 [−.15,.42]						
8. Enjoyment[2]	1.45	0.66	.01 [−.29,.31]	−.19 [−.47,.12]	−.21 [−.48,.09]	.48** [.21,.68]	.15 [−.15,.43]	−.02 [−.32,.28]	.06 [−.25,.35]					
Train Track Task														
9. Monitoring rate	4.28	2.17	−.18 [−.44,.11]	−.11 [−.39,.18]	.49** [.24,.68]	−.18 [−.45,.11]	.25 [−.05,.50]	.21 [−.09,.47]	.19 [−.11,.46]	−.08 [−.37,.22]				
10. Enjoyment[2]	1.40	0.57	.15 [−.14,.42]	.06 [−.23,.34]	−.22 [−.48,.07]	.33* [.04,.56]	−.18 [−.45,.12]	−.10 [−.39,.20]	−.25 [−.51,.05]	.23 [−.07,.50]	−.25 [−.49,.03]			
Wedgits Puzzle Task														
11. Monitoring rate	7.07	2.26	−.08 [−.36,.21]	−.08 [−.36,.22]	.34* [.06,.57]	−.26 [−.51,.03]	.1 [−.14,.44]	.1 [−.19,.40]	.11 [−.19,.40]	−.07 [−.36,.24]	.57** [.36,.73]	−.01 [−.29,.26]		
12. Enjoyment[2]	1.39	0.63	.28 [−.01,.53]	.03 [−.26,.32]	−.09 [−.37,.21]	.32* [.03,.56]	.20 [−.11,.47]	.11 [−.20,.40]	.07 [−.24,.36]	.14 [−.17,.42]	.25 [−.03,.49]	.10 [−.18,.36]	.12 [−.16,.38]	
Composite Monitoring Rate														
13. Monitoring rate (mean of TT and W)	5.67	1.98	−.14 [−.41,.15]	−.11 [−.38,.19]	.46** [.20,.66]	−.25 [−.50,.05]	.23 [−.07,.49]	.18 [−.12,.45]	.17 [−.14,.44]	−.08 [−.37,.23]	.88** [.80,.93]	−.14 [−.40,.14]	.89** [.82,.94]	.21 [−.07,.46]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95 % confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). RT = Reaction Time; eRT = estimated Reaction Time; CJ = Confidence Judgments. Monitoring rate = Monitoring behaviour per minute. * indicates $p < .05$. ** indicates $p < .01$. *N*s per cell ranged from 43 to 52. [1] scale: 0 = faster to 1 = slower; [2] scale: 1 = a lot of fun to 4 = no fun at all; [3] scale: 0 = very unsure to 1 = very sure.

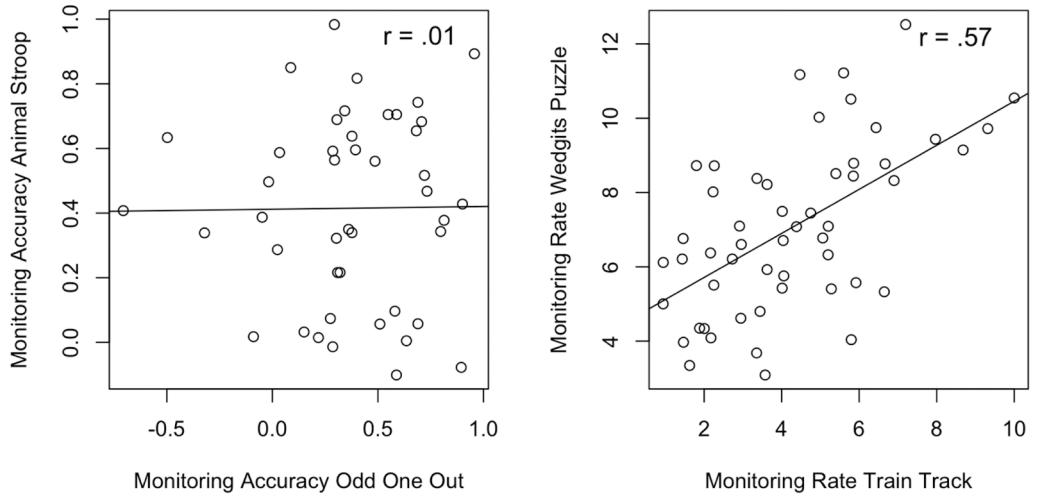

Fig. 4. Association of Monitoring Judgment Accuracies (left panel) and Monitoring Behaviours (right panel).

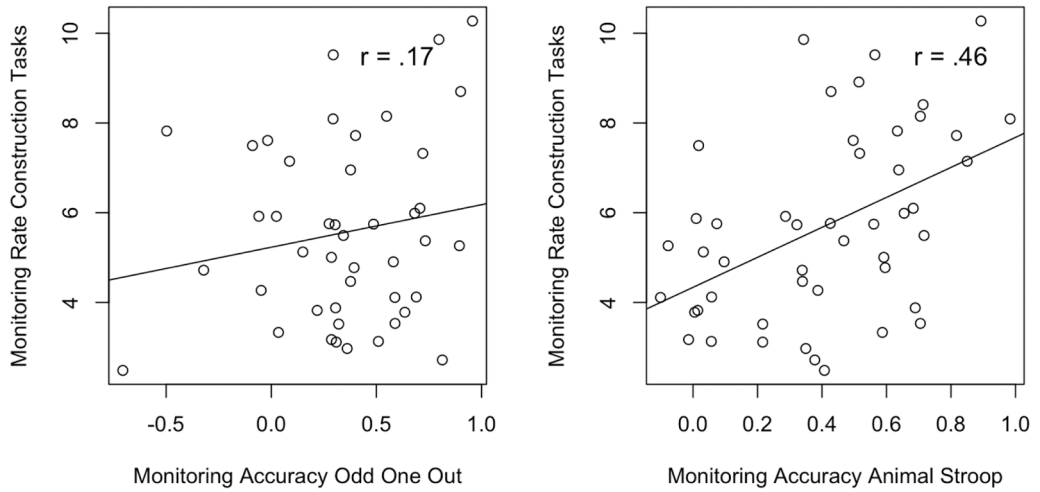

Fig. 5. Correlations between Monitoring Rate on Construction Tasks and Monitoring Accuracy on the Computer-based tasks (Left: Odd One Out, Right: Animal Stroop).

times) was positively associated with the frequency of monitoring behaviours evidenced during construction tasks. Overall, these findings lend support to and extend other studies that propose the construct of metacognitive monitoring is not unitary, that different processes probably underlie different types of monitoring and that the different indicators of monitoring should not be conflated without careful consideration of the processes involved.

4.1. Task dependency in monitoring judgment accuracy, task consistency in monitoring behaviours

The accuracy of children's monitoring judgments on the two computer-based tasks were not significantly correlated. That is, children who were more sensitive to trial-by-trial variations in their performance on one task were not more sensitive to variations in their performance on the other task. These results are in line with findings from adults (Dentakos et al., 2019; Kelemen et al., 2000; McDonough, Enam, Kraemer, Eakin, & Kim, 2021) and extend this line of research to a younger sample. In some ways the two tasks were similar – both were computerized, novel tasks, administered in a standardized way, and prompted judgments about performance were given after every trial via a visual analogue scale. Further, the monitoring accuracy scores of both tasks were significant and not different from one another, showing that on the group level children were similarly sensitive towards their performance in the tasks. However, in some important ways, the tasks also differed: the object of monitoring in the Odd One Out task is memory strength and in the Animal Stroop task it is reaction time. Koriat's (2012) cue utilization theory can account for this data pattern. This well supported theory states that various cues such as perceptual features, beliefs about a task, and familiarity with a task can contribute to monitoring judgments and that which cues contribute may vary depending on the primary task, the type of monitoring judgment, the person's

experience with the task etc. In the current study, it is likely that different cues are available during each task, for example recall fluency in the Odd One Out and the strength of the incorrect motor activation in the Animal Stroop. It should be acknowledged that while the adult literature on introspective reaction times shows that adults' judgments of reaction times are typically quite accurate (with the exception of complex dual-task contexts) and can be influenced by various sources of information (e.g., Bratzke & Bryce, 2022, 2023; Pavailler, Gevers, & Burle, 2024), it is not yet clear which cues may contribute to children's estimates of their reaction times or whether this trial-by-trial sensitivity translates to conscious awareness of what influences their performance. Further, children may have more experience with monitoring aspects of their memory than their reaction times. These factors may have led to a lack of association between monitoring accuracies. This would suggest that the information entailed in the primary task is responsible for the task specificity of monitoring judgments.

While the lack of an association between monitoring judgment accuracy across two tasks suggests a task dependency, the association between rates of unprompted monitoring behaviours suggests consistency across tasks. That means, children who showed a higher monitoring rate on one of the construction tasks, also showed a higher monitoring rate on the other construction task. A possible explanation could be that similar monitoring strategies (e.g., checking the plan) could be applied on both tasks despite the different task materials. Further, these strategies may be beneficial on other tasks and games that children had played previously, such as completing a jigsaw puzzle. Possibly, children who had already gained knowledge or experience that these strategies led to successful task completion on other tasks were able to generalise across tasks and apply them in our novel tasks. In turn, children who had not gained this experience or had it demonstrated for them might have fewer strategies available and therefore not shown these behaviours on either task.

It is notable that this finding stands in contrast to the findings of Spektor-Levy et al. (2017), the only other study that also assessed children's monitoring skills on two construction tasks, who found no significant associations. While similar coding schemes were applied in that study and the present one, in Spektor-Levy et al. (2017) fewer indicators of monitoring were considered and total counts of monitoring were analysed (whereas in the current study the rate of monitoring behaviours per minute was calculated). Further, the much larger age range included in that study may have contributed to the different results – it is conceivable that the association between monitoring behaviours across tasks changes with age. If so, a different pattern for older children might have driven the lack of association. Even though we observed intra-individual stability in the tendency to engage in monitoring behaviours, in the current study there is empirical evidence that different primary tasks elicit different levels of monitoring behaviour. That is, children showed a significantly higher monitoring rate on the Wedgits puzzle task than in the Train Track Task. On a descriptive level, Spektor-Levy et al. (2017) also found higher levels on one task (Train Track Task) than the other task (LEGO construction task). Future studies should specifically investigate possible developmental changes in the task stability of monitoring behaviour and employ a variety of primary tasks.

If supported by further empirical studies, the finding that the accuracy of monitoring judgments seems to be task-dependent and the tendency to engage in monitoring behaviours is more stable within individuals would have many implications both for research and application. As already acknowledged, metacognitive monitoring is a cognitive process that cannot be directly observed. Consequently, different measurement approaches have been developed to capture the construct and typically one indicator of monitoring is selected in a study. The current findings clarify that we should be cautious when generalising across studies that measure monitoring differently. While the tendency to engage in monitoring behaviours may be more generalisable and could be considered a trait (in this age range, at least), measures of monitoring judgment accuracy may be assessing the current state of metacognitive monitoring. As such, measures of monitoring behaviours may be more suited to investigations regarding individual differences, whereas monitoring judgments may be more appropriate for experimental designs. Further, the current findings highlight that how each type of monitoring is related to other constructs, such as academic achievement, may need to be revisited and clarified.

These findings may also provide insights into how to promote more effective monitoring in young children. To relate the two types of monitoring to real-life scenarios, these findings suggest that training or reminders to engage in monitoring behaviours such as checking one's own progress, self-questioning, and reviewing materials are likely to transfer to learning contexts outside of those in which they are trained. In contrast, when considering how to promote the accuracy of children's monitoring judgments, a blanket approach may be misguided and a more nuanced approach involving sensitive feedback and/or self-reflection may be required. For instance, it may be that when reflecting on how well one has mastered a new maths skill, speed of responding is a valid and useful cue to consider. In contrast, when reflecting on how well one has learned some new foreign language vocabulary, speed of responding may not be a valid cue and other cues more related to memory strength should rather be emphasised.

4.2. Partial links between monitoring judgments and monitoring behaviours

Another important aspect of the current study is that it examined the associations between the accuracy of children's monitoring judgments and their monitoring behaviours. Results indicated that children's monitoring accuracy regarding their reaction times (the Animal Stroop task), but not regarding their memory recall (the Odd One Out task), was significantly associated with their monitoring behaviour on the construction tasks. This partial link between monitoring behaviours and only one type of monitoring judgment was not expected and raises many interesting new questions. In any case, based on this data pattern it cannot be generally stated that the more one engages in monitoring behaviours, the more accurate one's monitoring judgments become. Instead, there appears to be a specificity regarding the types of monitoring judgments and the engagement in monitoring behaviour.

A possible explanation for this data pattern might be related to the required object of monitoring. The construction tasks as well as judgments of reaction times have in common that what is to be monitored is associated with externally observable behavioural responses (e.g., the mouse click on the animal on the screen, and the selection and combination of construction pieces), whereas

monitoring of memory recall requires the monitoring of internal, unobservable cognition (e.g., memory strength). This would predict that if monitoring behaviours were assessed during a task that makes greater demands on memory, thus necessitating more monitoring of internal cognitive processes, an association between the accuracy of memory monitoring judgments and monitoring behaviours may be observed.

When we consider more closely the cues that might be used in generating monitoring judgments of reaction times in the Animal Stroop task, it seems likely that feeling of conflict is an important and valid cue (see Desender, Van Opstal, & Van den Bussche, 2017 for related work on this in adults). That is, when the inhibition task induces cognitive conflict (because the larger animal on the screen is not the larger animal in real-life), this feeling of conflict is translated into a longer estimated reaction time. It could be that children who are attuned to this conflict are also more likely to be aware of a related feeling of conflict in the construction tasks, which elicits externally observable monitoring behaviours. Indeed, conflict can also be experienced in each of the construction tasks, for instance when one must dismantle and rebuild the Wedgits tower to ensure stability. Similar to the explanation for why performance on the Tower of Hanoi is primarily related to inhibition (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000), this detour away from a proximate goal in order to achieve task success elicits cognitive conflict (referred to as "goal-subgoal conflicts" in Miyake et al., 2000). While it is conceivable that this commonly experienced feeling of conflict is what drives the association between the accuracy of estimated reaction times and monitoring behaviour in the construction tasks, this is clearly a post hoc explanation deserving of further investigation. For instance, one would predict that estimated reaction times in a task that does not involve cognitive conflict would not be so strongly associated with monitoring behaviours in these tasks. Further, collecting feeling of conflict judgments from participants in these tasks may prove to be informative.

In summary, these novel findings regarding the associations between monitoring judgment accuracy and monitoring behaviours paint a complex and differentiated picture. Consistent with the finding that monitoring judgment accuracy is task dependent, it seems that associations between monitoring behaviours and monitoring judgments cannot be assumed and are influenced by specificities of the primary task. These findings have important consequences for the selection of monitoring indicators for example when evaluating interventions. One should consider carefully which aspect of monitoring is likely to be affected by the intervention activities and select the appropriate indicator.

4.3. Limitations and future direction

The current study is not without limitations. First, the tasks in which we collected monitoring judgments were purposefully designed to have similar features, as were the tasks in which we observed monitoring behaviours. However, one could argue that the two construction tasks were more similar than the two computer-based tasks, as they had similar task instructions (i.e., creating a construction based on a plan) and similar behaviours were coded (e.g., checking the plan). However, the tasks differed in terms of the type of material (train track vs. 3D puzzle pieces) and specific task demands (e.g., the Wedgits puzzle required construction on three dimensions, and the Train Track on only two dimensions). To address this concern, future studies could assess the accuracy of children's monitoring judgments in tasks that are more similar (e.g., two memory tasks with integrated confidence judgments) or monitoring behaviour could be assessed using tasks that differ more (e.g., a science experiment). As it stands, it cannot be determined whether the lack of association between prompted monitoring judgment accuracies is due to the different tasks or the different judgment types. Second, most participants of the current study had highly educated parents and they spoke German at home. Even though we do not have a strong hypothesis that children from other backgrounds would have performed differently on the tasks, this cannot be ruled out based on our data. Still, it is important to keep in mind that we applied within-child comparisons (i.e., children completed all of the tasks and we compared the differences between the tasks), meaning that any between-child differences such as family background or gender would have had a comparable effect on all the tasks. Thus, it is unlikely that sample characteristics had a strong influence on our overall conclusions. Third, we were particularly interested in comparing children's metacognitive monitoring across tasks that are typically used in the field to assess this skill. Therefore, we prompted children to provide monitoring judgments in two computer-based tasks, and we observed children's unprompted, naturally occurring monitoring behaviour in two construction tasks. To further unpack children's ability to report on their metacognitive monitoring in computer-based and construction tasks, in future studies children could also be prompted to provide monitoring judgments after each construction task. Fourth, we calculated the relative monitoring accuracy as our measure of children's monitoring judgment accuracy. However, it is possible that other ways of calculating monitoring accuracy (e.g., absolute monitoring accuracy) would have yielded different results. Fifth, while previous studies have also addressed similar questions with intercorrelations (e.g., Dentakos et al., 2019), it has been argued that simple correlational analyses are not ideal to test task dependency as it is not clear how strong the correlation needs to be to provide evidence for task stability (McDonough et al., 2021). A comparison of general mixed effect models could be a better data analytic approach. However, this type of analysis requires many trials per task which is a challenging endeavour for studies employing observational coding of monitoring behaviours. Finally, it should be acknowledged that we did not reach our target sample size based on our power analyses and that for certain analyses the sample size reduced further. This may have limited our power to detect correlations smaller than $r = .38$. Nevertheless, since the data pattern is quite unequivocal we do not consider this a threat to our main conclusions.

5. Conclusions

The current study was conducted with the aim of providing empirical evidence regarding the nature of metacognitive monitoring in young children, specifically focusing on the independence or relatedness of processes underlying monitoring judgments and monitoring behaviours. The data indicate that different processes underly these two types of monitoring and that they should not be

conflated. While children's tendency to engage in monitoring behaviours seems to be consistent across tasks, the accuracy of their monitoring judgments does not show the same intraindividual stability. These findings highlight the fact that metacognitive monitoring is a multifaceted construct and that the selection of monitoring indicators should be carefully considered. They inspire new ideas about how best to promote different aspects of monitoring and highlight the need for more targeted studies to uncover how the relationships among different aspects of monitoring may change over development.

Funding information

This project was funded by a LEAD intramural grant which was awarded to Janina Eberhart and Donna Bryce and a Deutsche Forschungsgemeinschaft grant awarded to Donna Bryce (BR 6057/3-1). This research was also partially supported by the Alexander von Humboldt Foundation (the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research) awarded to Kou Murayama.

CRediT authorship contribution statement

Bryce Donna: Writing – review & editing, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sakaki Michiko:** Writing – review & editing, Conceptualization. **Murayama Kou:** Writing – review & editing, Conceptualization. **Eberhart Janina:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

We would like to thank Selina Braun, Sarah Gläser, Greta Thunhorst, and Franziska Ingendahl for their support with study preparation, data collection, and video coding.

Author contribution

Janina Eberhart and Donna Bryce conceptualized and designed the study. Kou Murayama and Michiko Sakaki provided feedback during the conceptualization phase. Janina Eberhart and Donna Bryce prepared the study material. Janina Eberhart led the data collection process. Janina Eberhart and Donna Bryce led video coding. Janina Eberhart and Donna Bryce completed data analysis. Janina Eberhart wrote the first draft of the manuscript. Donna Bryce reviewed and edited previous versions. All authors commented on previous versions. All authors read and approved the final manuscript.

Data availability

Data will be made available on request.

References

Alloway, T.P. (2007). Automated Working Memory Assessment. Pearson Assessment.

Bratzke, D., & Bryce, D. (2022). Timing of internal processes: Investigating introspection about the costs of task switching and memory search. *Attention, Perception, & Psychophysics, 84*(5), 1501–1508. https://doi.org/10.3758/s13414-022-02510-6

Bratzke, D., & Bryce, D. (2023). Subjective estimates of total processing time in dual-tasking: (Some) good news for bad introspection. *Psychological Research, 87*(5), 1560–1568. https://doi.org/10.1007/s00426-022-01762-z

Bryce, D., Szűcs, D., Soltész, F., & Whitebread, D. (2011). The development of inhibitory control: An averaged and single-trial Lateralized Readiness Potential study. *NeuroImage, 57*(3), 671–685. https://doi.org/10.1016/j.neuroimage.2010.12.006

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem-solving. *Metacognition and Learning, 7*(3), 197–217. https://doi.org/10.1007/s11409-012-9091-2

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

de Abreu, P. M. J. E., Puglisi, M. L., Cruz-Santos, A., Befi-Lopes, D. M., & Martin, R. (2014). Effects of impoverished environmental conditions on working memory performance. *Memory, 22*(4), 323–331. https://doi.org/10.1080/09658211.2013.781186

Dentakos, S., Saoud, W., Ackerman, R., & Toplak, M. E. (2019). Does domain matter? Monitoring accuracy across domains. *Metacognition and Learning, 14*(3), 413–436. https://doi.org/10.1007/s11409-019-09198-4

Desender, K., Van Opstal, F., & Van den Bussche, E. (2017). Subjective experience of difficulty depends on multiple cues. *Scientific Reports, 7*, Article 44222. https://doi.org/10.1038/srep44222

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning, 3*(3), 189–206. https://doi.org/10.1007/s11409-008-9026-0

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

Eberhart, J., Ingendahl, F., & Bryce, D. (2024). Are metacognition interventions in young children effective? Evidence from a series of meta-analyses. *Metacognition and Learning, 20*(7). https://doi.org/10.1007/s11409-024-09405-x

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

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

Händel, M., & Dresel, M. (2022). Structure, relationship, and determinants of monitoring strategies and judgment accuracy. An integrated model and evidence from two studies. *Learning and Individual Differences, 100*, 1–13. https://doi.org/10.1016/j.lindif.2022.102229

Henry, L. A., Messer, D. J., & Nash, G. (2014). Testing for near and far transfer effects with a short, face-to-face adaptive working memory training intervention in typical children. *Infant and Child Development, 23*(1), 84–103. https://doi.org/10.1002/icd.1816

Janssen, N., & Lazonder, A. W. (2024). Meta-analysis of interventions for monitoring accuracy in problem solving. *Educational Psychology Review, 36*(3), 96. https://doi.org/10.1007/s10648-024-09936-4

Kälin, S., & Roebers, C. M. (2020). Time-based measures of monitoring in association with executive functions in kindergarten children. *Zeitschrift Für Psychologie, 228*(4), 244–253. https://doi.org/10.1027/2151-2604/a000422

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory Cognition, 28*(1), 92–107. https://doi.org/10.3758/BF03211579

Kim, Y., Zepeda, C. D., & Butler, A. C. (2023). An interdisciplinary review of self-regulation of learning: Bridging cognitive and educational psychology perspectives. *Educational Psychology Review, 35*(3), Article 92. https://doi.org/10.1007/s10648-023-09800-x

Koriat, A. (2012). The subjective confidence in one's knowledge and judgements: Some metatheoretical considerations. In In. M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition* (pp. 213–233). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199646739.003.0014.

Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13*(3), 441–453. https://doi.org/10.1111/j.1467-7687.2009.00907.x

Marulis, L. M., & Nelson, L. J. (2021). Metacognitive processes and associations to executive function and motivation during a problem-solving task in 3–5 year olds. *Metacognition and Learning, 16*(1), 207–231. https://doi.org/10.1007/s11409-020-09244-6

McDonough, I. M., Enam, T., Kraemer, K. R., Eakin, D. K., & Kim, M. (2021). Is there more to metamemory? An argument for two specialized monitoring abilities. *Psychonomic Bulletin Review, 28*(5), 1657–1667. https://doi.org/10.3758/s13423-021-01930-z

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory Cognition, 38*(4), 441–451. https://doi.org/10.3758/MC.38.4.441

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*(1), 19–46. https://doi.org/10.1007/s11409-013-9094-7

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (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 new findings. *Psychology of Learning and Motivation, 26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Pavailler, N., Gevers, W., & Burle, B. (2024). Temporal metacognition: Direct readout or mental construct? The case of introspective reaction time. Journal of Experimental Psychology: General. ⟨https://hal.science/hal-04753515⟩.

Peirce, J.W., Gray, J.R., Simpson, S., MacAskill, M.R., Höchstenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. Behavior Research Methods. http://doi.org/10.3758/s13428-018-01193-y.

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

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

Roebers, C. M., Cimeli, P., Röthlisberger, M., & Neuenschwander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: an explorative study on their interrelations and their role for school achievement. *Metacognition and Learning, 7*(3), 151–173. https://doi.org/10.1007/s11409-012-9089-9

Roebers, C. M., & Howie, P. (2003). Confidence judgments in event recall: developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85*(4), 352–371. https://doi.org/10.1016/S0022-0965(03)00076-6

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

Spektor-Levy, O., Basilio, M., Zachariou, A., & Whitebread, D. (2017). Young children's spontaneous manifestation of self-regulation and metacognition during constructional play tasks. *Teachers College Record, 119*(13), 1–28. https://doi.org/10.1177/016146811711901314

von der Linden, N., Löffler, E., & Schneider, W. (2016). Effects of a short strategy training on metacognitive monitoring across the life-span. *Frontline Learning Research, 3*(4), 37–55. https://doi.org/10.14786/flr.v3i4.196

Whitebread, D., Coltman, P., Pino Pasternak, D., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, Ds (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63–85. https://doi.org/10.1007/s11409-008-9033-1