

Testing the dependency of metacognitive monitoring and executive functions in children: a cross-sectional and longitudinal study

Donna Bryce, Janina Eberhart

Angaben zur Veröffentlichung / Publication details:

Bryce, Donna, and Janina Eberhart. 2026. "Testing the dependency of metacognitive monitoring and executive functions in children: a cross-sectional and longitudinal study." *Metacognition and Learning* 21(1): 49. <https://doi.org/10.1007/s11409-026-09497-7>.

Nutzungsbedingungen / Terms of use:

CC BY 4.0





Testing the dependency of metacognitive monitoring and executive functions in children: A cross-sectional and longitudinal study

Donna Bryce^{1,3} · Janina Eberhart^{2,3}

Received: 4 March 2026 / Accepted: 30 August 2026
© The Author(s) 2026

Abstract

Accurate metacognitive monitoring is crucial for effective regulation of learning and performance, yet little is known about its cognitive determinants in childhood. The present study investigated the role of executive functions – specifically inhibition and updating of working memory – in children’s metacognitive monitoring accuracy. While prior research has established correlational links between executive functions and metacognitive skills, the current experimental study aimed to provide insight into the causal nature of these relationships by systematically manipulating executive function demands within a task and examining the impact on local monitoring judgment accuracy. Sixty children from two age groups (7 and 9 years) participated, providing cross-sectional data; children in the younger cohort were reassessed approximately 17 months later, contributing longitudinal data. Data were analysed using linear mixed effect models and the results indicated that executive functions and metacognitive monitoring have a dependent relationship in childhood. Moreover, this dependency strengthened with age for inhibition and with repeated task exposure for both inhibition and working memory. These findings are consistent with a competition for resources account of the association between executive functions and metacognitive monitoring, suggesting that when cognitive resources are occupied resolving cognitive conflict or maintaining items in memory, fewer resources remain for monitoring processes and monitoring accuracy suffers. Results are also interpreted within the cue-utilization view of metacognitive monitoring, suggesting developmental changes in the cues children rely on when forming monitoring judgments and a differential dependency of these cues on executive functions.

Keywords Metacognition · Executive functions · Inhibition · Working memory · Cross-sectional · Longitudinal

✉ Donna Bryce
donna.bryce@uni-a.de

¹ Department of Psychology, University of Augsburg, Augsburg, Germany

² Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

³ Department of Psychology, University of Tübingen, Tübingen, Germany

Testing the dependency of metacognitive monitoring and executive functions in children: A cross-sectional and longitudinal study

It is well established that learners who make accurate judgments about their own understanding and cognition, and then sensitively adjust their behaviours or strategies based on this understanding experience more immediate task success (Bryce & Whitebread, 2012) and longer-term academic achievement (Dent & Koenka, 2016). While metacognitive monitoring processes seem to develop earlier than metacognitive control processes (Metcalf & Finn, 2013; Wan et al., 2025), the ability to accurately monitor oneself improves over the primary school years (van Loon et al., 2017) and shows meaningful individual differences until adulthood (Maki, 1998). In the current study we investigate potential determinants of accurate monitoring in young learners, namely the executive functions inhibition and updating of working memory. While previous studies have provided valuable evidence for an association between executive functions and metacognitive skills, the current experimental study aims to uncover novel insights into the causal nature of the relationships by manipulating the executive function demands of a task and examining the impact on metacognitive monitoring accuracy.

Metacognitive monitoring

Procedural metacognition describes behaviours and cognitive processes that serve to take control of and guide our own cognition while we complete a challenging task, and metacognitive monitoring is an important component of this (Flavell, 1979). Descriptively, monitoring is the process by which we self-observe our own learning, abilities, and performance during task processing. Successful metacognitive regulation relies on accurate monitoring – only if we are aware of dips in our performance or understanding can we make necessary adjustments to maximise our learning (such adjustments are driven by metacognitive control processes). Investigations into metacognitive monitoring often prompt participants to provide monitoring judgments on single items or trials (local judgments) or regarding a set of items or trials (global judgments; Schraw, 2009). These judgments can be predictions (e.g., judgments of learning, whereby a person predicts their likelihood of recalling an item in a test situation) or postdictions (e.g., confidence judgments, whereby a person reports their confidence regarding a response just given). Importantly, the precise monitoring judgment, and therefore the object of monitoring (Händel, Naujoks-Schober & Dresel, 2023), is selected based on what constitutes the most salient indicator of performance in a specific task, and can include memory trace strength, speed, and accuracy. The advantage of collecting metacognitive monitoring judgments is that these can be compared to the person's objective performance to calculate a measure of monitoring accuracy. Depending on the nature of the primary task and the type of monitoring judgment collected, various measures of monitoring accuracy can be calculated. These can be broadly categorised into absolute accuracy measures and relative accuracy measures (Schraw, 2009). Measures of absolute accuracy inform us about how close a monitoring judgment is to objective performance, whereas measures of relative monitoring accuracy inform us about how sensitive a person is to trial-by-trial fluctuations in their performance. Recognising fluctuations in our own performance might be particularly important for determining changes of strategy during a

task; for this reason, in the current study we focused on the relative accuracy of monitoring judgments.

There are two dominant and complementary models regarding how monitoring judgments are formed. The first, proposed by Nelson and Narens (1990) as a model of metamemory, posits that two levels are involved when we behave metacognitively during cognitive tasks: The object-level is the level of the cognitive activity, and the meta-level contains our representation of the activity and our own performance. Monitoring processes are defined as the flow of information from the object-level to the meta-level whereby this information updates our representation within the meta-level. Metacognitive control processes are defined as the flow of information from the meta-level to the object-level which asserts some change at the object-level. Nelson and Narens' conceptualisation of monitoring assumes that we have direct access to aspects of our cognitive processing or memory traces; this view is still reflected in more recent conceptualisations of monitoring processes (e.g., Li et al., 2021). The second dominant model of metacognitive monitoring, proposed and refined by Koriatic and colleagues, is called the Cue-Utilization Theory (Koriat, 1997). This takes a broader view of what might contribute to the accuracy of monitoring judgments and acknowledges additional sources of information that may be used as cues for these judgments, aside from information gained through direct access to cognitive processes. For instance, Koriat (2012) distinguishes between information-based and experience-based approaches or cues, the former of which may include pre-existing beliefs about the difficulty of specific types of items and the latter of which may include the experience of processing fluency. These two models are not to be understood as contradictory to one another; rather, Koriat's is consistent with Nelson and Narens model and adds a more nuanced perspective how monitoring judgments may be formed.

Executive functions

Individual differences in the accuracy of monitoring judgments exist (Maki, 1998) and understanding the reasons for these differences is of both practical and theoretical relevance. One potential candidate for supporting or influencing our ability to monitor ourselves accurately is the sophistication of our executive functions (EFs). EF is an umbrella term for a set of skills that allow us to react flexibly in the face of novel or challenging situations (Anderson, 2002). Typically, three EF components are investigated in adults, namely inhibition, updating of working memory and cognitive flexibility (Miyake et al., 2000). Inhibition is required to resist distraction from irrelevant information, to resolve cognitive conflict, and to halt inappropriate motor responses; working memory allows us to maintain information for a short time in memory and simultaneously manipulate it; cognitive flexibility is needed to switch between rules, tasks or perspectives (Miyake et al., 2000). While these three component EFs are identified in adults, data from children seem to better fit a single-factor structure (Willoughby et al., 2012) or a two-factor structure, whereby inhibition and cognitive flexibility are combined (Lee et al., 2013); for this reason, in the current study only two EFs are investigated. EFs develop rapidly between childhood and adulthood (Davidson et al., 2006) and are often seen as core cognitive skills that are called for in various complex tasks (Miyake et al., 2000). The fact that performance in EF tasks fluctuates daily (Wang et al., 2021; Blume et al., 2022) and can wane across multiple trials (Gerstadt et al., 1994)

leads to the characterisation of EFs as either themselves constituting a limited resource, or as drawing on cognitive resources that are limited. Given that EFs and metacognition have been associated with common brain regions and show similar developmental trajectories, it is not surprising that the association between metacognitive skills and EFs has long been considered of interest (e.g., Fernandez-Duque et al., 2000; Roebbers, 2017).

Metacognition and executive functions

Prior studies have provided evidence of positive associations between metacognitive skills and EFs in children (Bryce et al., 2015; Kälén & Roebbers, 2020, 2022; Marulis & Nelson, 2021; Roebbers et al., 2012; Dutemple et al., 2024; Theobald & Brod, 2025; Samara et al., 2023; Effeney et al., 2013; Bellon et al., 2019; Jiao et al., 2023). Some studies have also reported a lack of positive associations (Filippi et al., 2020; Destan & Roebbers, 2015; Tsalas et al., 2017; Spiess et al., 2016). In adults, more attention has been directed towards the relationship between monitoring and working memory (WM). For instance, adults with better WM monitor themselves more accurately in a separate task (Perrotin et al., 2007; Griffin et al., 2008; Boduroglu et al., 2014; Golke et al., 2022), and better monitoring accuracy about WM performance has been observed in younger than older adults (Touron et al., 2010) and in those with higher than lower WM capacity (Komori, 2016). Forsberg and colleagues (2021) extended these findings to include both children and adults – in their study, individuals with higher WM capacity more accurately judged their own WM performance. Across children and adults, these findings mostly emerge from correlational studies and data on metacognition and EFs are often collected with separate tasks that introduce additional, task-specific cognitive demands. Consequently, while these studies have provided a valuable evidence base suggesting a positive association between EFs and metacognition, they provide only limited insights into the nature of the relationship between metacognitive monitoring and EFs. The key question is, do metacognitive monitoring and EFs have a dependent relationship, or are they related for coincidental reasons? The answer to this question will have important consequences for how to promote these crucial cognitive skills and processes, with the aim of maximising gains in children's self-regulation and self-regulated learning.

To address issues of dependence or causality, intervention or experimental designs are required. Some such studies hint at a dependent relationship. For instance, children in an experimental group who were forced to “stop and think” before providing their monitoring judgments produced more accurate monitoring judgments (Wacker & Roebbers, 2022). In the same study, the usual association between inhibition skills and monitoring accuracy was observed, but notably this association was no longer present in the experimental group. These results are consistent with the idea that poor inhibition skills and the resulting impulsive responding negatively impact the accuracy of monitoring judgments. Individuals with better inhibition skills take the time to reflect more on their judgments and therefore provide more accurate judgments. In the opposite direction, recent meta-analytic results indicated that metacognition interventions lead to improvements in EF outcomes (Eberhart et al., 2025b), although it must be acknowledged that some of the interventions included in this analysis also targeted aspects of executive functioning.

In order to more directly test the proposition that one EF, the updating of WM, and metacognitive monitoring rely on the same limited cognitive resources, Bryce and colleagues

(2023) manipulated the WM demands within experiments and examined the impact of these manipulations on University students' relative monitoring accuracy. In two out of three experiments, relative monitoring accuracy suffered when WM demands increased, but not when other task difficulty demands increased (i.e., visual similarity, semantic complexity, and visual degradation). These results were interpreted as consistent with a competition for resources account of the relationship between WM and metacognitive monitoring. That is, since monitoring and WM rely on the same limited cognitive resources, when these resources are being used to maintain more items in WM, fewer resources are available for monitoring processes, and monitoring accuracy is reduced. Individuals with higher WM capacity monitor more effectively because maintaining items uses fewer resources for them. Similarly, within an individual, higher WM demands leave fewer resources for monitoring, leading to poorer monitoring performance. Findings consistent with Bryce et al. (2023) were also reported by Conte et al. (2023) in an experiment in which adult participants indicated their confidence in an n-back task. One prior study with children provided findings that are also consistent with the competition for resources account: Freeman and colleagues (2017) reported that 10-year-olds' monitoring accuracy suffered when the working memory demands of the task increased. In the current study we build on and extend this line of enquiry by investigating two EFs (inhibition and WM), two age groups of children (7- and 9-year-olds), and incorporating both a cross-sectional and longitudinal design.

The current study

In the current study we aimed to address the research question, what is the nature of the relationships between two EFs (inhibition and the updating of WM) and metacognitive monitoring and how stable are these relationships across childhood? Specifically, we test a competition for resources account of the relationship in children in a similar manner as done in the study by Bryce et al. (2023). In our study, child participants took part in two metacognitive EF experiments, one inhibition and one WM experiment, in which they provided monitoring judgments about their performance on each trial. Children were sampled from the first and third year of primary school in Germany (contributing cross-sectional data), and the children in the younger group were invited to participate again in the study approximately 17 months later (thus contributing longitudinal data). Based on prior findings, our first hypothesis (H1) was that monitoring accuracy will improve across development. Namely, monitoring judgments will be more accurate in the older than younger age group (cross-sectional analysis) and at the second testing time than the first (longitudinal analysis). Our second hypothesis (H2) was that EFs and monitoring are causally related in childhood when assessed experimentally. To test this, we manipulated the respective EF demands within each experiment and examined the impact of this manipulation on monitoring judgment accuracy. The core idea is, that if EFs and monitoring processes have a dependent relationship in children, monitoring accuracy should suffer when EF demands increase. Lastly, we tested the hypothesis that the relationship between EFs and monitoring will change across development (H3). We expected age-related changes in the relationships primarily because both metacognition and EFs develop rapidly in middle childhood. However, given the lack of related empirical evidence our H3 did not further specify the direction of change. Indeed, three result patterns were considered conceivable: The effect of EF

demands on monitoring accuracy could become stronger or weaken with age (these would be consistent with H3) or they could remain the same (which would be contrary to H3).

Methods

Preregistration, stimuli and data

This study was preregistered. The preregistration document, stimuli and data are shared via the dedicated Open Science Framework page <https://osf.io/v39nm/>.

Participants

A power analysis was conducted to estimate the sample sizes required to observe significant interaction effects in the cross-sectional analysis. For this, the G*Power program (Faul et al., 2007) was used; in particular, the statistical test ‘ANOVA: Repeated measures, within-between interaction’ (as recommended by A. Buchner in personal communication). Since this study represents a novel application of these experiments to children, effect sizes can only be estimated. In the adult version of the Metacognitive Odd One Out task (Bryce et al., 2023), a medium-to-large effect size of Sequence Length was observed (paired-samples t-test, Cohen’s $d=0.72$). As such, an a-priori power analysis was conducted based on a medium effect size of Cohen’s $f=0.25$, a significance level of 0.05 and a power of 0.9. According to these parameters, a minimum sample size of 23 in each age group was calculated to be sufficient. The fully within-subjects longitudinal design would have power to detect an even smaller effect. To account for likely participant attrition in the longitudinal design, we over-sampled the 6–7 year old group. It should be noted that in the end the data were analysed with linear mixed effects models (LMMs), meaning there is a discrepancy between our power analysis and the conducted analyses. We nevertheless took this decision due to the well-established advantages of LMMs, in particular that single trials are analysed instead of aggregated data, thereby increasing the amount of information used in the analysis. This approach generally provides greater statistical power and more precise estimates compared to analyses based on aggregated measures, especially in within-subject designs. Given this discrepancy and other limitations of power analyses, we treat the calculation made when planning the study as a transparent a priori estimate rather than definitive evidence of adequate power.

In total, 36 first graders and 26 third graders from different parts of Germany took part in a testing session. Two first graders chose to end testing session early, so their data are not included. Of the 34 first graders that remained (56% girls; $M_{age} = 7.28$ years; $SD_{age} = 0.34$), two did not complete one of our two tasks (the Metacognitive Odd One Out) so their data is missing from those analyses. All 26 third graders (42% girls; $M_{age} = 9.01$ years; $SD_{age} = 0.30$) provided complete data. All first graders were reintinvited to participate approximately 17 months later when they were in 3rd grade and 27 children (64% girls; $M_{age} = 8.55$ years; $SD_{age} = 0.35$) participated in this follow-up assessment. Time 2 data from the Metacognitive Odd One Out task was not used from the two participants who had not completed the task in the first testing session. Most of the children were from highly educated families as parents indicated that 81% had a university degree. 93% of children had German as their sole

native language, 5% were multilingual with German as one of their languages, and 2% had a language other than German as their native language; all children attended German schools and language presented no barriers during testing. Children were recruited through social media, flyers sent to sports clubs or family centres, and an email database of families who showed an interest in taking part in research studies. Children's parents provided consent and children gave verbal assent. This study was reviewed and approved by the Ethics Committee of the Faculty of Science at the University of Tübingen. Children received a 10€ gift card to thank them for their participation.

Measures

Metacognitive Odd One Out

The Odd One Out (adapted from Alloway, 2007) is a working memory task which has been used with children as young as 6 to 7 years old (de Abreu et al., 2014; Henry et al., 2014) and was adapted here to assess children's metacognitive monitoring. In this computerized task, children were shown three boxes on the computer screen with each box containing a shape (see Fig. 1a): two boxes always had the same shape, and one box had a different shape (odd one). Children were asked to click with the computer mouse in the box with the odd shape and to remember the location of the box. After a series of several odd shapes, children had to recall the order in which the odd shapes were presented by clicking in empty boxes on the screen. Then, children provided a confidence judgment on a continuous visual analogue scale (VAS) labelled with smileys regarding how confident they were that they recalled the location of the odd shapes correctly (this task has been previously reported in Eberhart et al. 2025a). Confidence regarding memory recall is the appropriate monitoring judgment in this task, as proportion correct recall is the primary indicator of performance.

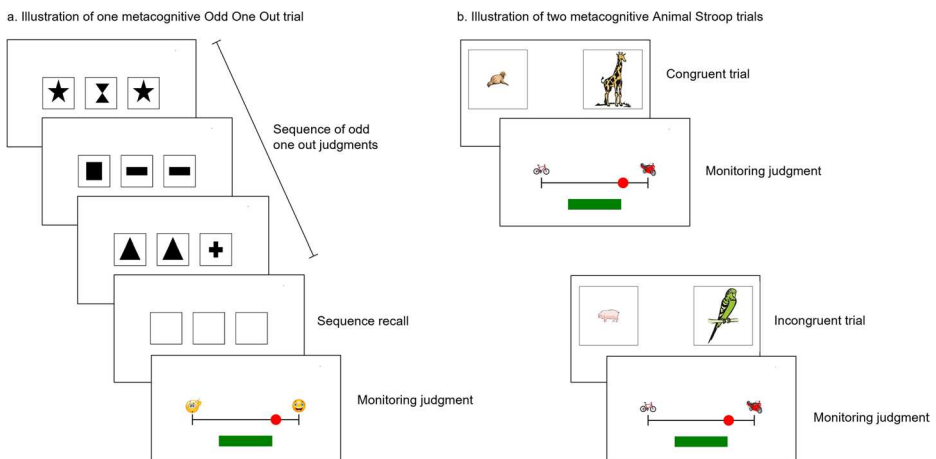


Fig. 1 Illustration of the Metacognitive Odd One Out (a; working memory) and the Metacognitive Animal Stroop (b; inhibition) Tasks. *Note:* a. The correct responses would be to click on the central box (screen 1), the left box (screen 2) and then the right box (screen 3), and then again repeat this sequence in the empty boxes in screen 4. b. The animals displayed in the congruent trial are a seal and a giraffe (correct response: giraffe), and in the incongruent trial are a pig and a budgerigar (correct response: pig)

Both the Metacognitive Odd One Out task and the Metacognitive Animal Stroop task (see below) were designed and generated using PsychoPy 3, version: 2021.2.0 (Peirce et al., 2019) and hosted online via www.pavlovvia.org (Bridges et al., 2020); children completed the tasks on their home PCs under the guidance of a research assistant who led the data collection session via online video conference. To ensure scalability across various screen sizes, item sizes are defined using height units. This method allows stimuli to automatically adjust based on the browser window's dimensions. For example, a height unit value of 0.4 means the stimulus will take up 40% of the total window height. 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. To provide their confidence judgment they could click anywhere on the VAS (with the slider granularity set to 0) and 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.

The task was then introduced in three steps. First, the research assistant explained that children had to identify the odd shape and to memorize the location in which the odd shape was presented. Children practiced the identification of the odd shapes and recall of the locations in three practice trials with a sequence length of one, two, and three respectively. The task was programmed such that in the identification phase, the screen only moved on if the children clicked on the odd shapes. When children made mistakes in recalling the locations in more than one practice trial, research assistants could add another practice trial. Second, children were introduced to the confidence judgments and the VAS, labelled with two smileys, following an approach developed by Roebbers et al. (2012). The research assistant explained that the children should click on the VAS towards the happy smiley to express that they were very confident about their answer and to click towards the puzzled smiley to express that they were not so confident. Then, children were asked to answer three questions where it was expected that children would vary on their confidence level and then indicate their confidence on the scale ("how old are you?", "how much hair do you have on your head?", "how old am I [the research assistant]?"). If necessary, instructions about the use of the VAS to provide a confidence judgment was repeated. Third, children completed three practice trials including the identification of the odd shapes, the recall of locations and the confidence judgments. The experimental trials consisted of 24 trials with a break incorporated after 12 trials. For motivational purposes, four trials with a sequence length of two were included. The other 20 trials had a sequence length of four (easy) and five (difficult) for the younger age group, and five (easy) and six (difficult) for the older age group; trials of each sequence length were presented in random order. For the participants participating in the longitudinal study, sequence lengths of five (easy) and six (difficult) were used at the second testing time. Sequence length was determined based on a piloting where a proportional recall accuracy of 0.7 was aimed for.

Metacognitive Animal Stroop

The Animal Stroop task is an inhibition task which was developed for use with children of this age range as well as adults (Bryce et al., 2011). In this task children were presented two pictures of animals on a screen next to each other with one of the animals presented larger on the screen than the other. Children were instructed to click as fast as possible on the

animal that would be larger in real life. This task includes congruent trials where the animal that would be larger in real life is also presented larger on the screen, and incongruent trials where the animal that would be larger in real life is presented smaller on the screen. The stimulus presentation size is irrelevant to task success and needs to be inhibited. A simple version of this task (without monitoring judgments) is shared online (Bryce, 2022). In the metacognitive version of the task, also used in Eberhart et al. (2025a), children were asked after each trial to indicate how fast their response was (estimated Reaction Time=eRT) on a continuous VAS labelled with a bicycle and motorcycle (see Fig. 1b). Although eRTs are less common in educational and developmental research, they have been used extensively in cognitive psychology as measures of introspective access to response latency (e.g., Corrallo et al., 2008). Prior work shows that eRTs are sensitive to experimental manipulations affecting objective reaction times (e.g., Bratzke & Bryce, 2022; Bratzke & Bryce, 2023), while also reflecting the use of non-temporal cues (e.g., Bryce & Bratzke, 2014), similar to other metacognitive monitoring judgments. We considered eRTs particularly appropriate because reaction time is the primary indicator of performance in the Animal Stroop task and therefore a salient target of monitoring, and they have been successfully used previously in this age range (Eberhart et al. 2025a). Furthermore, research on the development of time perception suggests that children can make reasonably accurate explicit temporal judgments, with temporal sensitivity approaching adult levels by approximately 8–9 years of age (Droit-Volet, 2016).

Similar to the Odd One Out task, this task was programmed in PsychoPy (Peirce et al., 2019) and hosted on Pavlovia (Bridges et al., 2020). A total of 36 animal images were used in this task, consisting of 12 small-sized, 12 medium-sized, and 12 large-sized animals based on their real-life dimensions. In each trial, two animals of different real-life sizes were randomly chosen from the set of images and displayed on the screen, each enclosed within a black frame. One animal image was presented smaller on the screen (15% of the height of the computer screen) and the other animal image was presented larger on the screen (40% of the height of the computer screen).

To familiarise the child with the various demands of this task, three practice blocks were administered. First, children identified which animal would be larger in real life in one congruent (giraffe and butterfly) and one incongruent (bear and ladybird) trial. Next, the VAS to provide the monitoring judgment was explained to the children. They were informed that all their responses would be fast, but if they thought they were very fast, they should click towards the motorcycle and if they thought they were not so fast they should click towards the bicycle. They then practiced estimating their reaction time with the help of three maze puzzles. These were presented on the screen and children moved through the maze using their computer mouse and clicked on an animal at the end. The three mazes had a different number of blockages, meaning that the time needed to complete them (and therefore, the child's reaction time) varied. After completing each maze, the child indicated their reaction time on the VAS and the research assistant could repeat the instructions about the VAS if necessary. Lastly, the child completed four Animal Stroop trials with the monitoring judgment included. After successfully completing the practice blocks, the child completed 48 experimental trials, with a break incorporated every 12 trials. Half the trials were congruent (representing low inhibition demands), and half were incongruent (representing high inhibition demands), presented in random order.

Procedure

Children completed both tasks within one online data collection session led by an experimenter who was present via video call. Children used their own home PCs (tablets and cell phones were not permitted) and were required to use a computer mouse during the session. Before the online testing session, informed consent and demographic data were gathered through an online survey conducted via the SoSciSurvey platform (Leiner, 2019). At the start of the video call, after verbally assenting to taking part in the session, the child shared their screen with the experimenter. The child practiced using the computer mouse by clicking on images of familiar cartoon characters which disappeared when clicked. Within the session, the two tasks were administered in a counterbalanced order. With the experimenter's assistance, the child completed the example and practice trials described above. Once the experimental trials commenced, the experimenter turned off their camera and microphone to minimize potential distractions. The entire testing session took no longer than 1h.

Data analysis

Data exclusion

A child's whole dataset for a task was only excluded if they did not complete that task during the testing session (already reported in the Participants section). Some trials in the Metacognitive Animal Stroop task had extremely long reaction times, for example if the child accidentally clicked outside of the window in which the experiment was running. Since these long reaction times do not reflect the child's cognitive processing time, it was decided these should not contribute to the calculation of children's monitoring accuracy. To this end, single trials with extremely long reaction times (calculated individually for each participant dataset: those that were more than two standard deviations above the participant mean reaction time) were excluded before data analyses (cross-sectional data: 4.1% of all trials; longitudinal data: 5.0% of all trials). It should be noted that this exclusion criterion was not pre-registered but was nevertheless deemed conservative and justified. No single trials were excluded from the Metacognitive Odd One Out task.

Planned analysis

Data were analysed in R (R Core Team, 2019) using the packages tidyverse (Wickham et al., 2019), lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017), sjPlot (Lüdtke, 2024) and plots produced using ggplot2 (Wickham, 2016). First, children's overall performance on each task as well as the effect of WM Load (in the Odd One Out task) or Congruency (in the Animal Stroop task) on their performance were analysed to establish whether the experimental manipulations functioned as expected. Hypothesis 1, that monitoring accuracy will improve across development, was addressed here – for each task, each participant's relative monitoring accuracy was calculated as the Pearson correlation between their Performance and Monitoring Judgment and these values were tested for differences between groups (cross-sectional data) or timepoints (longitudinal data).

To address the remaining hypotheses (H2 and H3), single trials were analysed in four linear mixed effects models (LMMs) – one for each task (Odd One Out and Animal Stroop) and each dataset (cross-sectional and longitudinal). In the following, the term EF demands will be used as a factor name. In the Odd One Out task this factor refers to the WM Load and has two levels: short and long sequence length; in the Animal Stroop task this factor reflects Congruency and has two levels: congruent and incongruent. Likewise, the term Performance will be used as a variable name. In the Odd One Out task this refers to the proportion of the sequence correctly recalled (Recall Accuracy); in the Animal Stroop task this refers to the reaction time. Finally, the term Monitoring Judgment will be used in place of CJ (in the Odd One Out) and eRT (in the Animal Stroop). In each of the four models Monitoring Judgment was the dependent variable, Performance, EF Demands, and Age Group (for the cross-sectional analyses) or Testing Time (for the longitudinal analyses) were entered as fixed effects. To aid interpretation, the Performance fixed effect was centred before being entered into the LMMs. For models analysing cross-sectional data, a random intercept for Participant was included to account for individual differences in baseline Monitoring Judgments, acknowledging that repeated trials within the same participant are not independent. For models analysing longitudinal data, the hierarchical structure of the data was modelled by including random intercepts for participants and for the interaction between participants and testing times.

Relative monitoring accuracy is defined as the relationship between monitoring judgments and performance; as such a significant effect of Performance within a LMM indicates monitoring accuracy (and is the equivalent of a significant correlation in typical resolution measures). An interaction of another factor with Performance, e.g., a Performance $\times$ EF Demands interaction, indicates that the extent to which Performance contributes to monitoring judgments is affected by this factor (e.g., EF Demands). In other words, monitoring accuracy is affected by EF Demands. First, the best fitting LMM was established for each model following the model selection procedure described by Barr and colleagues (2013). The full models include all predictors, their two-way interactions and their three-way interactions as fixed effects. In an iterative fashion, the impact of removing one term is tested using likelihood ratio tests and the best fitting LMMs are presented in tables. Additionally for each of the four models, two confirmatory analyses were conducted. To test whether EFs and monitoring are related in childhood when assessed experimentally (H2), a model containing the fixed effects of Performance, EF Demands and Age Group / Testing Time (M1) was compared with a model containing these effects as well as the Performance $\times$ EF Demands interaction (M2). To test whether the relationship between EFs and monitoring changes across development (H3), a model containing the fixed effects of Performance, EF Demands and Age Group / Testing Time as well as the theoretically important Performance $\times$ EF Demands and Performance $\times$ Age Group / Testing Time two-way interactions (M1) was compared with a model containing these effects as well as the Performance $\times$ EF Demands $\times$ Age Group / Testing Time three-way interaction (M2).

Results

The Results section is structured first by dataset (cross-sectional data, then longitudinal data) and then by task (Odd One Out, then Animal Stroop). Within this structure, the descriptive statistics and manipulation check for each task is presented first, and H1 (developmental changes in monitoring accuracy) will be addressed within these sections. Hypotheses 2 and 3 will be addressed in the proceeding sections where each LMM is presented. It is important to note that four sets of analyses are conducted to address each hypothesis (one for each EF and for each dataset), and the overall conclusions will be summarised in the Discussion.

Cross-sectional analyses

Descriptive statistics

Descriptive statistics for the cross-sectional data can be found in Table 1. In the Odd One Out task, all participants identified the odd shape with very high accuracy (ranging from 97% to 100%). Despite efforts to match the difficulty via pilot testing, the younger group performed significantly better than the older group in terms of proportional recall accuracy; this difference was also reflected in monitoring judgments with younger children giving higher CJs than older children. Younger children also took significantly longer to provide their CJs than older children. Interestingly, the groups were able to monitor their performance equally well (contrary to H1), with a moderate to high mean monitoring accuracy which was significantly different from zero for both groups (younger: $t(31)=13.47$, $p < .001$; older: $t(25)=15.66$, $p < .001$). This indicates children are sensitive to trial-by-trial

Table 1 Descriptive statistics for each task for the cross-sectional data

Variables from the OOO	Younger		Older		Group Δ
	<i>N</i>	Mean (SD)	<i>N</i>	Mean (SD)	<i>t</i>
Prop. odd shapes identified	32	0.99 (0.01)	26	0.99 (0.01)	-0.66, $p=.512$
Recall accuracy	32	0.74 (0.14)	26	0.66 (0.13)	2.20, $p=.032$
Confidence judgment (CJ)	32	0.74 (0.14)	26	0.64 (0.18)	2.43, $p=.018$
Time to give CJ (sec)	32	5.12 (2.05)	26	3.50 (0.90)	4.04, $p<.001$
Monitoring accuracy (r)	32	0.55 (0.23)	26	0.50 (0.16)	0.94, $p=.352$
<i>Variables from the AS</i>	Younger		Older		Group Δ
	<i>N</i>	Mean (SD)	<i>N</i>	Mean (SD)	<i>t</i>
Accuracy	34	0.96 (0.07)	26	0.94 (0.04)	0.90, $p=.370$
Reaction time (sec)	34	2.34 (0.71)	26	1.45 (0.26)	6.77, $p<.001$
Estimated reaction time (eRT)	34	0.18 (0.15)	26	0.18 (0.13)	-0.08, $p=.934$
Time to give eRT (sec)	34	4.54 (1.47)	26	3.28 (0.87)	4.14, $p<.001$
Monitoring accuracy (r)	34	0.29 (0.27)	26	0.34 (0.24)	-0.76, $p=.453$

OOO Odd One Out task (metacognitive working memory task), AS Animal Stroop task (metacognitive inhibition task)

The proportion of odd shapes identified, the recall accuracy, the CJ and the eRT variables could range from 0 to 1; the monitoring accuracy variables could range from -1 to 1

Two children from the younger group did not complete the Odd One Out task. The test for group differences reported in the last column is an independent samples t-test

variations in their WM performance. Further, one-way ANOVAs with the within-subjects factor WM Load were conducted for each age group on Recall Accuracy and CJ. These indicated that both groups recalled a lower proportion of locations in the high than the low WM Load condition (younger: $F(1, 31)=29.17, p < .001, \eta_p^2=0.48$; older: $F(1, 25)=28.04, p < .001, \eta_p^2=0.53$) and that this was reflected in their CJs (younger: $F(1, 31)=38.52, p < .001, \eta_p^2=0.55$; older: $F(1, 25)=50.01, p < .001, \eta_p^2=0.67$).

In the Animal Stroop task, the high mean accuracy indicates participants could process the task to a high standard. While older participants responded to the animal stimuli and gave their eRTs more quickly than younger participants, their mean eRTs were strikingly similar. Likewise, mean monitoring accuracy was very similar across groups (contrary to H1) and significantly different from zero (younger: $t(33)=6.31, p < .001$; older: $t(25)=7.44, p < .001$) indicating children are sensitive to trial-by-trial variations in their reaction times. As an additional manipulation check, one-way ANOVAs with the within-subjects factor Congruency were conducted for each age group on reaction time and eRT. These indicated that both groups were slower to respond to incongruent than congruent trials (younger: $F(1, 33)=23.66, p < .001, \eta_p^2=0.42$; older: $F(1, 25)=35.48, p < .001, \eta_p^2=0.59$) and that this was reflected in their eRTs (younger: $F(1, 33)=13.72, p = .001, \eta_p^2=0.29$; older: $F(1, 25)=14.06, p = .001, \eta_p^2=0.36$).

Working memory: Monitoring judgment accuracy

The best fitting LMM for the cross-sectional data from the Odd One Out task is presented in Table 2 and the data pattern can be viewed in Fig. 2a. As in all upcoming LMMs, the dependent variable was the monitoring judgment (in this case, the CJ) and a significant contribution of Performance (in this case, Recall accuracy) reflects relative monitoring accuracy. Overall, the model explained a reasonable proportion of variance in CJs, with the fixed effects explaining 24% of variance (Marginal R^2) and the random effects (or, differences between individuals) explaining an additional 25% of variance (Conditional $R^2=0.49$). The best fitting model indicates that children's recall accuracy significantly positively predicted their CJs (i.e., they demonstrated significant monitoring accuracy), and that WM Load significantly negatively predicted their CJs (i.e., they gave lower CJs in the high than the low WM Load condition). Additionally, confirmatory analyses were conducted to address H2 and H3. First, a likelihood ratio test indicated that including the Recall Accuracy $\times$ WM Load interaction term did not significantly improve the model, $\chi^2(1)=3.09, p = .079$, indicating that WM Load (i.e., the EF demands of the task) did not affect monitoring accuracy (contrary to H2). Second, a likelihood ratio test indicated that including the three-way inter-

Table 2 Estimates for fixed effects in the best fitting LMM for the cross-sectional data on the Odd One Out task

Predictor	Estimate	95% CI	<i>p</i>
(Intercept)	0.75	0.70–0.79	<0.001
Recall accuracy	0.41	0.37–0.46	<0.001
WM Load	-0.11	-0.13 – -0.08	<0.001

LMM Linear Mixed-Effects model, CI Confidence interval

Confidence judgment is the dependent variable, *p*-values are based on the Wald statistic

Recall accuracy was centred. WM load was dummy coded with Low WM load as reference level

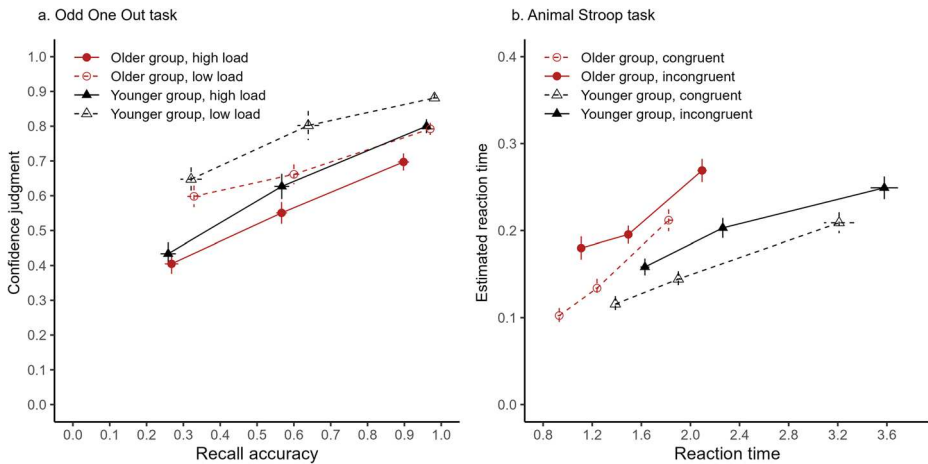


Fig. 2 Results from the Cross-Sectional Analyses of the Odd One Out task (**a**) and the Animal Stroop Task (**b**). *Note.* In both plots, Monitoring Judgment is plotted against Performance. For illustration single trials were vincentized based on Performance (single trial data was analysed). That is, the first vincentile in each panel represents 33% of trials with the lowest performance for each participant by EF Demand condition; the second contains the next 33% of trials with lowest performance, and so on. Error bars represent ± 1 within-subject *SE*

action (Group $\times$ Recall Accuracy $\times$ WM Load) did not significantly improve the model, $\chi^2(1)=0.05$, $p = .820$, reflecting that the nature of the relationship between WM load (or, EF demands) and monitoring accuracy was not different for the two age groups (contrary to H3).

Inhibition: Monitoring judgment accuracy

The best fitting LMM for the cross-sectional data from the Animal Stroop task is presented in Table 3. Overall, it should be noted that only a very small proportion of the variance in eRTs was accounted for by the fixed effects in this model (9.4%), with 43% of variance being attributed to the random effects. Additionally, two confirmatory analyses were con-

Table 3 Estimates for fixed effects in the best fitting LMM for the cross-sectional data on the Animal Stroop task

Predictor	Estimate	95% CI	<i>p</i>
(Intercept)	0.18	0.14–0.22	<0.001
Reaction time	0.04	0.03–0.05	<0.001
Congruency	0.03	0.02–0.04	<0.001
Reaction time $\times$ Group	0.07	0.05–0.09	<0.001
Reaction time $\times$ Congruency	-0.01	-0.02–0.00	0.205
Reaction time $\times$ Congruency $\times$ Group _{Older}	-0.05	-0.08 – -0.02	0.002

LMM Linear Mixed-Effects model, *CI* Confidence interval

Estimated reaction time is the dependent variable, *p*-values are based on the Wald statistic

Recall accuracy was centred. Congruency and Group were dummy coded with Congruent and Younger group as reference levels

ducted to address the hypotheses. First, a likelihood ratio test indicated that including the Reaction Time $\times$.

Congruency (i.e., the EF demands of the task) interaction significantly improved the model, $\chi^2(1)=9.47$, $p = .002$. The associated negative estimate indicates that reaction time contributes less to estimated reaction time in incongruent than congruent trials, in other words monitoring accuracy suffers in incongruent trials or when EF demands increase (consistent with H2). The other planned model comparison indicated that including the three-way interaction (Reaction Time $\times$ Congruency $\times$ Group) significantly improved the model, $\chi^2(1)=10.31$, $p = .001$, reflecting that the nature of the relationship between Congruency (or, EF demands) and monitoring accuracy changed across the age groups (consistent with H3). A further examination of the estimates in Table 3 shows that the effect of congruency on monitoring accuracy is greater in the older than in the younger group¹. This is reflected in the two red lines having different gradients, and the two black lines being rather parallel in Fig. 2b.

Longitudinal analyses

Descriptive statistics

Descriptive statistics for the longitudinal data can be found in Table 4. As in the cross-sectional data, children whose data entered the longitudinal analyses had no problems identifying the odd shapes in the Odd One Out task (identification rates ranged from 97% to 100%). Although recall accuracy was well matched in this dataset, we still see that when they are younger, children give more positive confidence judgments and take a little longer to provide their monitoring judgments than when they are older. Monitoring accuracy was again moderate to high and significantly different from zero (Time 1: $t(24)=11.63$, $p < .001$; Time 2: $t(24)=15.52$, $p < .001$) and there was no difference in monitoring accuracy from time 1 to time 2 (contrary to H1). At both time points the WM Load affected children's recall accuracy (Time 1: $F(1, 24)=36.65$, $p < .001$, $\eta_p^2=0.60$; Time 2: $F(1, 24)=16.32$, $p < .001$, $\eta_p^2=0.40$) as well as their CJs (Time 1: $F(1, 24)=31.73$, $p < .001$, $\eta_p^2=0.57$; Time 2: $F(1, 24)=14.49$, $p = .001$, $\eta_p^2=0.38$).

Also in the longitudinal dataset, mean monitoring accuracy in the Animal Stroop task was significantly different from zero at both time points (Time 1: $t(26)=5.34$, $p < .001$; Time 2: $t(26)=8.64$, $p < .001$) and not statistically different at time 1 and time 2 (contrary to H1). One-way ANOVAs with the within-subjects factor Congruency for each testing time indicated that both groups were slower to respond to incongruent than congruent trials (Time 1: $F(1, 26)=18.06$, $p < .001$, $\eta_p^2=0.41$; Time 2: $F(1, 26)=6.10$, $p = .020$, $\eta_p^2=0.19$) and that this was reflected in their eRTs (Time 1: $F(1, 26)=8.93$, $p = .006$, $\eta_p^2=0.26$; Time 2: $F(1, 26)=14.69$, $p = .001$, $\eta_p^2=0.36$).

¹ To check whether the effects were driven by age-related differences in speed of processing, additional analyses were conducted where each child's mean reaction time in congruent trials was included as a covariate (fixed effect) in the LMMs. The pattern of results was not altered.

Table 4 Descriptive statistics for each task for the longitudinal data

Variables from OOO	Time 1		Time 2		Time Δ
	<i>N</i>	Mean (SD)	<i>N</i>	Mean (SD)	<i>t</i>
Prop. odd shapes identified	25	0.99 (0.01)	25	0.99 (0.01)	-1.74, $p=.094$
Recall accuracy	25	0.76 (0.12)	25	0.72 (0.14)	1.33, $p=.195$
Confidence judgment (CJ)	25	0.75 (0.14)	25	0.61 (0.16)	4.11, $p<.001$
Time to give CJ (sec)	25	5.25 (2.27)	25	4.34 (1.28)	2.38, $p=.025$
Monitoring accuracy (<i>r</i>)	25	0.56 (0.24)	25	0.55 (0.18)	0.16, $p=.874$
<i>Variables from AS</i>	Time 1		Time 2		Time Δ
	<i>N</i>	Mean (SD)	<i>N</i>	Mean (SD)	<i>t</i>
Accuracy	27	0.97 (0.03)	27	0.95 (0.05)	1.66, $p=.110$
Reaction time (sec)	27	2.42 (0.74)	27	1.72 (0.44)	5.45, $p<.001$
Estimated reaction time (eRT)	27	0.18 (0.15)	27	0.23 (0.11)	-1.47, $p=.154$
Time to give eRT (sec)	27	4.64 (1.60)	27	3.85 (1.09)	2.69, $p=.012$
Monitoring accuracy (<i>r</i>)	27	0.30 (0.29)	27	0.40 (0.24)	-1.54, $p=.137$

OOO Odd One Out task (metacognitive working memory task), AS Animal Stroop task (metacognitive inhibition task)

The proportion of odd shapes identified, the recall accuracy, the CJ and the eRT variables could range from 0 to 1; the monitoring accuracy variables could range from -1 to 1

Two children did not complete the Odd One Out task at Time 1. The test for group differences reported in the last column is a paired-samples *t*-test

Working memory: Monitoring judgment accuracy

The best fitting LMM for the longitudinal data on the Odd One Out task is presented in Table 5 and the data pattern can be seen in Fig. 3a. For this model, Marginal $R^2 = 0.32$ and Conditional $R^2 = 0.45$. Please note, because WM load and Time were dummy coded with Low WM load and T1 as reference levels, the model estimates effects as deviations from this baseline condition. Consequently, only interaction terms comparing other WM load $\times$ Time combinations to the reference level appear in the fixed-effects table; the effect of recall accuracy at Low WM load and T1 is captured by the main effect of recall accuracy. As in previous analyses, confirmatory tests were conducted to address hypotheses two and three.

Table 5 Estimates for fixed effects in the best fitting LMM for the longitudinal data on the Odd One Out task

Predictor	Estimate	95% CI	<i>p</i>
(Intercept)	0.78	0.72–0.83	<0.001
Recall accuracy	0.46	0.36–0.56	<0.001
WM Load	-0.08	-0.11 – -0.04	<0.001
Time	-0.12	-0.18 – -0.06	<0.001
Recall accuracy $\times$ WM Load _{Low} $\times$ Time _{T1}	-0.03	-0.17–0.11	0.660
Recall accuracy $\times$ WM Load _{High} $\times$ Time _{T1}	0.03	-0.11–0.16	0.700
Recall accuracy $\times$ WM Load _{Low} $\times$ Time _{T2}	0.15	0.01–0.29	0.033

LMM Linear Mixed-Effects model, CI Confidence interval

Confidence judgment is the dependent variable, *p*-values are based on the Wald statistic

Recall accuracy was centred. WM load and Time were dummy coded with Low WM load and T1 as reference levels

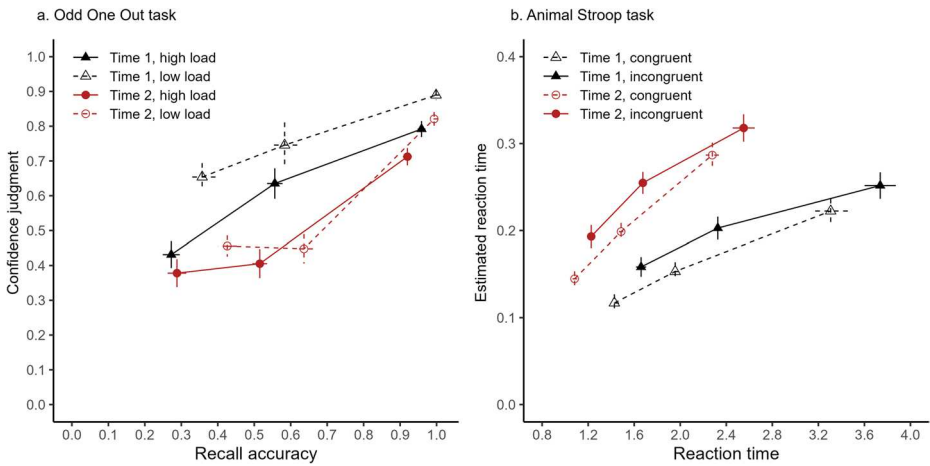


Fig. 3 Results from the Longitudinal Analyses of the Odd One Out task (a) and the Animal Stroop Task (b). *Note.* In both plots, Monitoring Judgment is plotted against Performance. For illustration single trials were vincentized based on Performance (single trial data was analysed). That is, the first vincentile in each panel represents 33% of trials with the lowest performance for each participant by EF Demand condition; the second contains the next 33% of trials with lowest performance, and so on. Error bars represent ± 1 within-subject *SE*

Regarding H2, a likelihood ratio test indicated that including the Recall Accuracy $\times$ WM Load interaction term did not significantly improve the model, $\chi^2(1)=0.86$, $p = .353$, indicating that collapsed across testing times WM Load (i.e., the EF demands of the task) did not impact the extent to which recall accuracy contributed to the monitoring judgment (in other words, WM Load did not affect monitoring accuracy; partially contrary to H2). However, the nature of the relationship between WM load and monitoring accuracy appeared to change across testing times (consistent with H3), since the comparison of models with and without the three-way interaction (Time $\times$ Recall Accuracy $\times$ WM Load) indicated the three-way interaction improved the model, $\chi^2(1)=4.37$, $p = .037$. When we examine Table 5 it becomes clear that at Time 1, monitoring accuracy was unaffected by WM load, whereas at Time 2 monitoring accuracy suffered when WM load increased. To further illustrate this, the mean correlation coefficients for each condition were: $r_{T1, LowWM}=.46$, $r_{T1, HighWM}=.53$, $r_{T2, LowWM}=.64$, $r_{T2, HighWM}=.48$. This result indicates that when encountering the metacognitive Odd One Out task a second time, participants showed a similar data pattern as adults did in Bryce et al. (2023) – their monitoring accuracy suffered when WM load (or, EF demands) increased.

Inhibition: Monitoring judgment accuracy

For the Animal Stroop task, the results of the longitudinal data analyses largely mirrored the cross-sectional results. The best fitting LMM for the longitudinal data on the Animal Stroop task is presented in Table 6 and the data are depicted in Fig. 3b. It is again notable that only a small proportion of the variance in eRTs was accounted for by the model, with Marginal $R^2=0.09$ and Conditional $R^2=0.44$. Regarding H2, the likelihood ratio test indicated that including the Reaction Time $\times$ Congruency interaction significantly improved the

Table 6 Estimates for fixed effects in the best fitting LMM for the longitudinal data on the Animal Stroop task

Predictor	Estimate	95% CI	<i>p</i>
(Intercept)	0.16	0.12–0.20	<0.001
Time	0.09	0.02–0.15	0.009
Reaction time	0.04	0.03–0.05	<0.001
Congruency	0.02	0.01–0.03	0.003
Reaction time × Time	0.05	0.03–0.07	<0.001
Reaction time × Congruency	-0.01	-0.02–0.00	0.262
Reaction time × Congruency × Time _{T2}	-0.04	-0.06 – -0.02	0.001

LMM Linear Mixed-Effects model, *CI* Confidence interval

Estimated reaction time is the dependent variable, *p*-values are based on the Wald statistic

Recall accuracy was centred. Congruency and Time were dummy coded with Congruent and T1 as reference levels

model, $\chi^2(1)=9.34$ $p = .002$ and a negative estimate indicated that monitoring accuracy is worse in incongruent than congruent trials (i.e., when EF demands increase, consistent with H2). Regarding H3, the three-way interaction (Reaction Time × Congruency × Time) also significantly improved the model, $\chi^2(1)=11.09$, $p < .001$. As in the cross-sectional data, the effect of congruency on monitoring accuracy is more pronounced at Time 2 than Time 1 (consistent with H3)².

Discussion

The primary aim of the present study was to uncover whether metacognitive monitoring and EFs have a dependent relationship in childhood, and whether the nature of the relationship changes with age. To this end, we conducted an experimental study designed to test how two EFs – inhibition and the updating of WM – are related to the accuracy of children’s monitoring judgments. The implementation of a mixed cross-sectional and longitudinal design provided a multifaceted view on the issue. Our findings indicate that metacognitive monitoring and EFs do indeed have a dependent relationship in children, and that this dependency gets stronger with age (inhibition) and with repeated exposure to the task (WM and inhibition). In the following, key results addressing each of the hypotheses will be summarised.

Based on prior developmental findings, we hypothesised (H1) that the accuracy of monitoring judgments would improve with age (cross-sectional analysis) and time (longitudinal analysis). Contrary to this, while children’s monitoring judgment accuracies were significantly different from zero in all analyses, they did not significantly change with age or time. One possible explanation is that our sample covered a relatively narrow age range, limiting the developmental variability available to detect age-related improvements. In addition, much of the existing evidence for developmental improvements in monitoring accuracy comes from studies using school-related learning materials, such as spelling or concept-learning tasks (e.g., van Loon et al., 2017). In contrast, our study used more abstract cognitive tasks, which may rely on somewhat different processes and may show weaker or slower developmental changes in monitoring accuracy. It is therefore possible that developmental

² To check whether the effects were driven by age-related differences in speed of processing, additional analyses were conducted where each child’s mean reaction time in congruent trials was included as a covariate (fixed effect) in the LMMs. The pattern of results was not altered.

improvements in monitoring are more readily observed in educational contexts that children encounter frequently and for which they receive regular external feedback. Furthermore, the longitudinal interval may have been too short to capture subtle developmental changes. Thus, rather than indicating an absence of development in monitoring, our findings may suggest that such changes are task-dependent and/or emerge over longer developmental periods. Descriptively, there was some evidence that children were better at monitoring their performance in the WM task than in the inhibition task. This finding was also reflected in the results of Eberhart et al. (2025a) and is possibly due to having more practice reflecting on the strength of their memory traces than on estimating response times. Indeed, this finding is consistent with mounting evidence suggesting that monitoring judgments and their accuracy are very task- and context-dependent (e.g., Mengelkamp & Bannert, 2010; McDonough et al., 2021; Kelemen et al., 2000; Dentakos et al., 2019). The descriptive statistics additionally provided important evidence that children understood all aspects of the tasks and can monitor their performance in classic EF tasks, that our manipulations of EF demand functioned as expected, and that there were no ceiling or floor effects in performance in either of the primary tasks. The latter is especially important when assessing relative monitoring accuracy, since variance in performance is necessary when evaluating how sensitive people are to changes in their performance. These features of the dataset speak to the validity of the tasks employed and their appropriateness for testing the second and third hypotheses.

Our second hypothesis was that EFs and monitoring are causally related in childhood, and the third hypothesis was that the relationship between EFs and monitoring changes across development (although we were not more specific about the direction of change). The analyses conducted to test H2 and H3 paint an interesting picture regarding how EFs and metacognitive monitoring might be related across childhood. Since the result patterns differed across the two EFs, we will consider them separately. The cross-sectional data indicated that on first encountering the metacognitive Odd One Out task, WM load did not affect the accuracy of children's monitoring judgments in either the younger or the older age group. This contrasts with the results of Bryce et al. (2023; specifically Experiment 1) whereby adults' monitoring accuracy suffered when WM load increased and could hint that children generate their monitoring judgments in a qualitatively different way than adults. However, it is important to highlight that the longitudinal data provided more nuanced insights. Namely, when children processed the task for a second time 17 months later, they made more accurate monitoring judgments in lower than higher WM load trials (i.e., a pattern consistent with the adult findings). Taken together, this data pattern suggests that not only age but also exposure to the task (albeit very limited exposure) affects the nature of the relationship between monitoring and WM. Our best explanation for this calls on the cue-utilization view of monitoring judgments and reflects the changing cues that children may use when forming monitoring judgments. Specifically, it seems that on first encountering this working memory task, children did not spontaneously compose their monitoring judgments in the same way as adults did, because their monitoring judgments did not suffer when their cognitive resources were 'used up' by higher WM demands. Children may base their monitoring judgments more on theory-based cues (i.e., their prior knowledge and beliefs about a task) than on experience-based cues (i.e., the immediate feelings that arise during task processing). We propose that increased WM demands primarily hinder the direct access to experiences during cognitive processing. However, with repeated experience with a task, children may begin generating their monitoring judgments similarly to adults as their judg-

ment accuracy is affected by the WM demands. This interpretation would be consistent with Koriati (1997) who found a move from theory-based to experience-based cues with practice.

In contrast to the nuanced findings regarding WM, the data patterns regarding the association between inhibition and monitoring were consistent across cross-sectional and longitudinal datasets. In addressing H2 with regards to inhibition, based on our data we can conclude that inhibition and monitoring have a dependent relationship in children because children monitored their performance more accurately when inhibition demands were lower (congruent compared to incongruent trials). With regards to H3, the data suggest that the nature of the relationship changes across development, because the effect of inhibition demands on monitoring accuracy was stronger in the older than the younger group of children (cross-sectional data) and when the same children were older and encountering the task for a second time (longitudinal data). The reason for the association becoming stronger with age may also reflect the changing nature of children's monitoring judgments, similar to the explanation elaborated in the previous paragraph. These data are consistent with other experimental findings that took a different approach to testing this relationship, namely using a "stop and think" instruction (Wacker & Roebbers, 2022).

Overall, the current findings demonstrate that competition for resources is likely one reason that EFs and monitoring judgment accuracy are associated in children (with the important caveat for WM, that this may only hold when children base their judgments on more experience-based cues or the strength of their memory trace). This may partially explain differences in monitoring accuracy on the one hand between individuals, and on the other hand within individuals across different contexts. That is, individuals with more sophisticated EFs are likely to make more accurate monitoring judgments than those with weaker EFs because more cognitive resources remain available for metacognitive monitoring processes. Also, within individuals we could expect monitoring judgments to be less accurate in novel or challenging contexts that require flexible responding (i.e., those that are highly demanding of EFs) compared to contexts when our cognitive processing can unfold more routinely or automatically. Importantly, due to its experimental design, these findings support a specific direction of causation, namely that changes in executive function demands affect monitoring processes.

Nevertheless, these findings do not exclude the possibility that a reverse direction of causation is also at play. That is, it is conceivable that individuals who more effectively monitor themselves and make more accurate monitoring judgments can more accurately recognise the need to engage and upregulate their EFs. This possibility could be tested in an experimental design in which monitoring is somehow prevented and the impact on EF performance is examined. Alternatively, training or intervention studies in which accurate monitoring is promoted and the impact on EF performance is evaluated could shed light on this possibility.

Further, the current findings do not exclude the possibility that EFs and monitoring processes are additionally associated for other reasons. Importantly, the current study focuses on only one operationalisation of metacognitive monitoring – the relative monitoring accuracy of local monitoring judgments. Whether increasing EF demands would similarly limit participants' engagement with monitoring behaviours or strategies remains an unanswered question. Further, prompted monitoring judgments (as collected here) may not reflect the way in which spontaneous monitoring judgments occur (Halamish & Undorf, 2020). Indeed, recent empirical findings from adults suggest that the occurrence of spontaneous monitoring

judgments is more closely related to the quality of monitoring strategies than the accuracy of prompted monitoring judgments (Händel et al., 2025). An interesting future direction of research could be to investigate how different aspects of metacognitive monitoring are supported by or dependent on EFs.

Given the different pattern of results across the two tasks with regards to how monitoring accuracy was affected by the manipulation of EF demands, it is important to consider whether this reflects a meaningful difference in the nature of the relationships between monitoring and each EF. One interpretation that could be drawn, is that both inhibition and WM compete for resources with monitoring, but inhibition does so more consistently and earlier in childhood than WM. However, we would caution against such an interpretation, because there are many other differences between the two tasks that should be considered. Most strikingly, the fact that children provided different types of monitoring judgments in the two tasks (confidence judgments and estimated reaction times) makes it difficult to draw strong conclusions about whether different result patterns are due to the EFs examined or the judgment type. Indeed, this was not one of our aims in designing this study, and it is a challenging issue to solve. It is important in any metacognitive task that the monitoring judgment provided reflects the performance measure of interest; in the case of WM and inhibition these tend to be different. To be able to compare across EFs, targeted studies are needed.

Limitations

As well as the above-mentioned confound of EF component and judgment type, there are some other limitations in the current study that should be acknowledged. First, despite selecting the sequence lengths based on pilot data, in the cross-sectional data we did have a performance difference in the metacognitive Odd One Out task, whereby the younger group recalled a higher proportion of the locations than the older group. A solution for future studies would be to make the task adaptive and ensure that a recall accuracy of approximately 70% is achieved for each individual. Secondly, one nice design feature of the experiments in Bryce and colleagues (2023) that was not implemented here is the inclusion of an additional ‘task difficulty’ manipulation. In the adult study, these additional manipulations had no impact on monitoring judgment accuracy, and as such it was confirmed that specifically increasing the WM demands affected monitoring processes. In the current study it was decided not to include such an additional manipulation, as this would have increased the task durations and made the data collection sessions unacceptably long for our child participants.

Our sample is also limited in certain aspects. That is, the current study only covers a small period of development and we can not comment on the nature of the relationships between EFs and monitoring before the age of 7 and between the ages of 9 years and adulthood. It would certainly be feasible to extend this type of investigation to older children and adolescents. Unfortunately, extending it to younger children could prove challenging, as different methods are more appropriate for assessing EF and metacognition in younger children. Lastly, the representativeness of our sample could be called into question. Despite our efforts to recruit via community groups and social media, the majority of children who took part had highly educated parents. Perhaps the requirements regarding computer equipment and a quiet testing environment inadvertently excluded a wider range of families from participating. To ensure a generalisable and representative developmental science, alternative methods of recruitment and data collection should be explored.

Practical implications

Although the current study was conducted to address very specific questions about the nature of the relationship between two sets of cognitive skills – EF and metacognition – some practical implications can also be drawn from these findings. First, since individual differences in executive functioning seem to contribute to the accuracy of monitoring processes, if teachers and instructors reduce the EF demands of tasks in which monitoring accuracy is particularly important, they should expect to see general metacognitive benefits. Second, since we know that our capacity to perform well on EF tasks fluctuates within individuals across time points, we could recommend that learners themselves reduce the demands on their EFs (e.g., via cognitive offloading or restricting the influence of irrelevant information) when they are low in cognitive resources or when monitoring accuracy is particularly important (e.g., while studying for a test). Lastly, the findings specific to WM and monitoring suggest that practice making monitoring judgments in the same task combined with individualized feedback may support children to use experience-based cues in their judgments which are sometimes considered more valid (Koriat et al., 2013), which may in turn best support children's developing monitoring skills.

Conclusions

The question of whether EFs and metacognition are related has long received attention (Fernandez-Duque et al., 2000) and has fairly consistently been answered in the affirmative by empirical evidence over the last 20 years (for a review: Roebers, 2017). The current study built on correlational findings and extended these with an experimental approach, aiming to understand the nature of the relationship between metacognitive monitoring and EFs in children, specifically testing if the relationship is dependent in nature. As such, the current study provided valuable novel insights into why two EFs – inhibition and the updating of WM – are associated with the accuracy of monitoring judgments. The results indicate that EFs and metacognitive monitoring do indeed have a dependent relationship in children, and the analysis of data collected both cross-sectionally and longitudinally provided a differentiated view on how the relationship changes in childhood. Inhibition and monitoring processes seem to compete for resources, since monitoring accuracy suffered when inhibition demands increased and this effect increased with age and time. In contrast, repeated exposure to the task seems to be required for WM to compete for resources with monitoring processes in children. These findings provide evidence for one pathway by which EFs and metacognitive monitoring are associated, which can inspire future studies to test additional hypotheses. Further, the finding that EFs and monitoring are not only coincidentally related, but appear to have a causal relationship in children, could inform the design of interventions to promote these crucial cognitive skills that underly children's ability to regulate themselves and their learning.

Acknowledgements We thank Celine Spannagl, Greta Thunhorst and Franziska Ingendahl for their support with recruitment and data collection. This study was supported by the Deutsche Forschungsgemeinschaft (BR 6057/3–1).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anderson, P. (2002). Assessment and Development of Executive Function (EF) During Childhood. *Child Neuropsychology*, 8(2), 71–82. <https://doi.org/10.1076/chin.8.2.71.8724>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bellon, E., Fias, W., & De Smedt, B. (2019). More than number sense: The additional role of executive functions and metacognition in arithmetic. *Journal of Experimental Child Psychology*, 182, 38–60. <https://doi.org/10.1016/j.jecp.2019.01.012>
- Blume, F., Irmer, A., Dirk, J., & Schmiedek, F. (2022). Day-to-day variation in students' academic success: The role of self-regulation, working memory, and achievement goals. *Developmental Science*, 25(6), e13301. <https://doi.org/10.1111/desc.13301>
- Boduroglu, A., Tekcan, A. I., & Kapucu, A. (2014). The relationship between executive functions, episodic feeling-of-knowing and confidence judgements. *Journal of Cognitive Psychology*, 26(3), 333–345. <https://doi.org/10.1080/20445911.2014.891596>
- 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 Psychologische Forschung*, 87(5), 1560–1568. <https://doi.org/10.1007/s00426-022-01762-z>
- Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8, e9414. <https://doi.org/10.7717/peerj.9414>
- Bryce, D. (2022). AnimalStroop - A child-friendly conflict task. *Pavlovian*. <https://doi.org/10.17605/OSF.IO/F48QP>
- Bryce, D., & Bratzke, D. (2014). Introspective reports of reaction times in dual-tasks reflect experienced difficulty rather than timing of cognitive processes. *Consciousness and Cognition*, 27, 254–267. <https://doi.org/10.1016/j.concog.2014.05.011>
- 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>
- 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., & Szűcs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition and Learning*, 10(2), 181–198. <https://doi.org/10.1007/s11409-014-9120-4>
- Bryce, D., Kattner, F., Birngruber, T., & Wellingerhof, P. (2023). Monitoring accuracy suffers when working memory demands increase: Evidence of a dependent relationship. *Journal of Experimental Psychology: Learning Memory and Cognition*, 49(12), 1909. <https://doi.org/10.1037/xlm0001262>

- Conte, N., Fairfield, B., Padulo, C., & Pelegrina, S. (2023). Metacognition in working memory: Confidence judgments during an n-back task. *Consciousness and Cognition*, 111, 103522. <https://doi.org/10.1016/j.concog.2023.103522>
- Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science*, 19, 1110–1117. <https://doi.org/10.1111/j.1467-9280.2008.02211.x>
- Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44(11), 2037–2078. <https://doi.org/10.1016/j.neuropsychologia.2006.02.006>
- 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 (Hove, England)*, 22(4), 323–331. <https://doi.org/10.1080/09658211.2013.781186>
- Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(3), 425–474. <https://doi.org/10.1007/s10648-015-9320-8>
- 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>
- Destan, N., & Roebbers, 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>
- Droit-Volet, S. (2016). Development of time. *Current Opinion in Behavioural Sciences*, 8, 102–109. <https://doi.org/10.1016/j.cobeha.2016.02.003>
- Dutemple, E., Brokl, C., & Poulin-Dubois, D. (2024). I think therefore I learn: metacognition is a better predictor of school readiness than executive functions. *Frontiers in Developmental Psychology*, 2, 1332358. <https://doi.org/10.3389/fdyps.2024.1332358>
- Eberhart, J., Ingendahl, F., & Bryce, D. (2025b). Are metacognition interventions in young children effective? Evidence from a series of meta-analyses. *Metacognition and Learning*, 20(1), 7. <https://doi.org/10.1007/s11409-024-09405-x>
- Eberhart, J., Murayama, K., Sakaki, M., & Bryce, D. (2025a). Metacognitive monitoring in early elementary school-aged children: Task dependency in monitoring judgments, task consistency in monitoring behaviours. *Cognitive Development*, 74, 101561. <https://doi.org/10.1016/j.cogdev.2025.101561>
- Effeney, G., Carroll, A., & Bahr, N. (2013). Self-regulated learning and executive function: Exploring the relationships in a sample of adolescent males. *Educational Psychology*, 33(7), 773–796. <https://doi.org/10.1080/01443410.2013.785054>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition*, 9(2), 288–307. <https://doi.org/10.1006/ccog.2000.0447>
- Filippi, R., Ceccolini, A., Periche-Tomas, E., & Bright, P. (2020). Developmental trajectories of metacognitive processing and executive function from childhood to older age. *Quarterly Journal of Experimental Psychology*, 73(11), 1757–1773. <https://doi.org/10.1177/1747021820931096>
- 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>
- Forsberg, A., Blume, C. L., & Cowan, N. (2021). The development of metacognitive accuracy in working memory across childhood. *Developmental Psychology*, 57(8), 1297. <https://doi.org/10.1037/dev0001213>
- Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, 57, 58–64. <https://doi.org/10.1016/j.lindif.2017.06.003>
- Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 312–7 years old on a stroop-like day-night test. *Cognition*, 53(2), 129–153. [https://doi.org/10.1016/0010-0277\(94\)90068-X](https://doi.org/10.1016/0010-0277(94)90068-X)
- Golke, S., Steininger, T., & Wittwer, J. (2022). What makes learners overestimate their text comprehension? The impact of learner characteristics on judgment bias. *Educational Psychology Review*, 34(4), 2405–2450. <https://doi.org/10.1007/s10648-022-09687-0>
- Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, 36(1), 93–103. <https://doi.org/10.3758/Mc.36.1.93>

- Händel, M., Nett, U., Bryce, D., & Dresel, M. (2025). Unraveling the reciprocal effects and micro-development of metacognitive monitoring strategies and monitoring judgments: A micro-genetic diary study. *Learning and Individual Differences*, 122, 102748. <https://doi.org/10.1016/j.lindif.2025.102748>
- Händel, M., Naujoks-Schober, N., & Dresel, M. (2023). Metacognitive Monitoring via Strategies and Judgments: Different Phases, Different Objects. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, 55(2–3), 67–76. <https://doi.org/10.1026/0049-8637/a000273>
- Halamish, V., & Undorf, M. (2020). Do learners spontaneously monitor their memory? The gap between educational-developmental and cognitive-experimental research. *Zeitschrift für Psychologie*, 228(4), 304. <https://doi.org/10.1027/2151-2604/a000429>
- 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>
- Jiao, X., Zhang, A., & Bu, X. (2023). Relating metacognition and executive functions to early mathematical and language skills in children aged 4–5 years. *Metacognition and Learning*, 18(2), 449–464. <https://doi.org/10.1007/s11409-023-09337-y>
- Kälin, S., & Roebbers, 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>
- Kälin, S., & Roebbers, C. M. (2022). Longitudinal associations between executive functions and metacognitive monitoring in 5-to 8-year-olds. *Metacognition and Learning*, 17(3), 1079–1095. <https://doi.org/10.1007/s11409-022-09306-x>
- 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>
- Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of group differences using a listening span test. *Frontiers in Psychology*, 7, 285. <https://doi.org/10.3389/fpsyg.2016.00285>
- Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2013). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 117–135). Psychology Press.
- 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. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A. (2012). The subjective confidence in one's knowledge and judgements: Some metatheoretical considerations. 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>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13> target="blank"><https://doi.org/10.18637/jss.v082.i13>><https://doi.org/10.18637/jss.v082.i13>targettype="DOI" class="RefTarget">
- Lee, K., Bull, R., & Ho, R. M. (2013). Developmental changes in executive functioning. *Child Development*, 84(6), 1933–1953. <https://doi.org/10.1111/cdev.12096>
- Leiner, D. J. (2019). *Sosci survey* (Version 3.1.06). <https://www.soscisurvey.de>
- Li, H. H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *Neuron*, 109(22), 3699–3712. <https://doi.org/10.1016/j.neuron.2021.08.022>
- Lüdtke, D. (2024). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.17, <https://CRAN.R-project.org/package=sjPlot>
- Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 131–158). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410602350-13>
- 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](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Perrotin, A., Belleville, S., & Isingrini, M. (2007). Metamemory monitoring in mild cognitive impairment: Evidence of a less accurate episodic feeling-of-knowing. *Neuropsychologia*, 45(12), 2811–2826. <https://doi.org/10.1016/j.neuropsychologia.2007.05.003>
- R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Roebbers, 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>
- Roebbers, 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>
- Samara, E., Kolovelonis, A., & Goudas, M. (2023). Examining the associations between calibration accuracy and executive functions in physical education. *European Journal of Educational Research*, 12(1). <https://doi.org/10.12973/eu-jer.12.1.359>
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, 4(1), 33–45. <https://doi.org/10.1007/s11409-008-9031-3>
- Spieß, M. A., Meier, B., & Roebbers, C. M. (2016). Development and longitudinal relationships between children's executive functions, prospective memory, and metacognition. *Cognitive Development*, 38, 99–113. <https://doi.org/10.1016/j.cogdev.2016.02.003>
- Alloway, T. P. (2007). *Automated Working Memory Assessment*. Pearson Assessment.
- Theobald, M., & Brod, G. (2025). The role of feedback and working memory for goal-related monitoring and goal revision. *Learning and Instruction*, 97, 102108. <https://doi.org/10.1016/j.learninstruc.2025.102108>
- Touron, D. R., Oransky, N., Meier, M. E., & Hines, J. C. (2010). Metacognitive monitoring and strategic behaviour in working memory performance. *Quarterly Journal of Experimental Psychology*, 63(8), 1533–1551. <https://doi.org/10.1080/17470210903418937>
- Tsalas, N., Sodian, B., & Paulus, M. (2017). Correlates of metacognitive control in 10-year old children and adults. *Metacognition and Learning*, 12(3), 297–314. <https://doi.org/10.1007/s11409-016-9168-4>
- van Loon, M., de Bruin, A., Leppink, J., & Roebbers, 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>
- Wacker, S., & Roebbers, C. M. (2022). Stop and think: Additional time supports monitoring processes in young children. *Plos One*, 17(9), e0274460. <https://doi.org/10.1371/journal.pone.0274460>
- Wan, Q., Savic, O., Gao, M., Ralston, R., O'Leary, A. P., & Sloutsky, V. M. (2025). From uncertainty to performance optimization: Longitudinal insights into metacognitive development. *Child Development*, 96(6), 2097–2111. <https://doi.org/10.1111/cdev.70029>
- Wang, Y., Zhang, Y., Chen, M., & Chen, Y. (2021). Adolescents' daily executive function: Methodological considerations, daily variation, and associations with daily experiences. *Psychological Assessment*, 33(10), 973. <https://doi.org/10.1037/pas0001040>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.
- Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*, 24(1), 226. <https://doi.org/10.1037/a0025361>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of OpenSource Software*, 4(43), 1686. <https://doi.org/10.21105/jos.s.01686>