Metacognition and Learning
https://doi.org/10.1007/s11409-024-09400-2

Development of metacognitive monitoring and control skills in elementary school: a latent profile approach

Mariëtte van Loon[1] · Claudia M. Roebers[2]

Received: 7 December 2023 / Accepted: 31 July 2024

Abstract

This study aims to understand individual differences between children in metacognitive monitoring and control processes and the developmental trajectories of metacognition over one year. Three indicators of procedural metacognition were used: monitoring accuracy (discrimination of confidence judgments between correct and incorrect test responses), effective restudy selections, and accuracy of response maintenance/withdrawal decisions. These indicators were measured for two tasks (text comprehension and Kanji memory) at two measurement points one year apart. Participants were 151 second graders (*M* age 7.61 years) and 176 fourth graders (*M* age 9.62 years). With latent profile analyses, distinct metacognition profiles were found for both grade levels at both measurement points. Children showed heterogeneity in the proficiency of metacognition but also in the extent to which metacognitive skills were generalizable across the two tasks. For second-grade children, being low at metacognition at the first measurement point was not associated with extra risks for low metacognition one year later. However, for fourth graders, children with low metacognitive skills appeared likely to stay low in metacognition over time and particularly showed ineffective restudy decisions. This indicates that they seemed at risk for a longer-term metacognitive deficiency. Findings may improve understanding of the heterogeneity of metacognition and support distinguishing typical from at-risk metacognitive development.

Keywords Monitoring · Control · Metacognition · Development · Elementary school · Latent profile analysis

Metacognition predicts performance in many learning domains for children and adults (Ohtani & Hisasaka, 2018). The definition of metacognition as “cognition about cognition” (Flavell, 1979) encompasses declarative and procedural components. Declarative metacog-

✉ Mariëtte van Loon
mariette.vanloon@psychologie.uzh.ch

[1] Department of Psychology, University of Zurich, Binzmühlestrasse 14, Zürich CH-8050, Switzerland

[2] University of Bern, Bern, Switzerland

Published online: 24 August 2024

nition involves knowledge of and beliefs about the factors that affect learning and cognition, whereas procedural metacognition involves skills to monitor and control one's learning while working on tasks. Procedural metacognition is strongly associated with the outcomes of self-regulated learning (Ohtani & Hisasaka, 2018). The Nelson and Narens (1990, 1994) model of metacognition explains procedural metacognition. Two levels of cognition, the meta-level and the object-level, are assumed to be interrelated via monitoring and control. To assess monitoring, a person can, for instance, be asked to make a judgment of learning (JOL) when studying an item or make a confidence judgment (CJ) about the correctness of their test answers. Measures of control are, for example, study time allocation, restudy selections, and the extent to which a person strategically decides to submit correct responses and opts out by withdrawing incorrect responses (Dunlosky et al., 2016; Koriat & Goldsmith, 1996; Krebs & Roebers, 2010). Research on procedural metacognition typically used single tasks and measures instead of combining multiple components. The present study combines multiple measures of procedural metacognition to obtain insights into the heterogeneity between elementary school children's procedural metacognition.

Consider, for example, Anna, a fourth-grade student, who has to read a text about bear hibernation and solve math questions. She needs to apply procedural metacognition to understand the text and the math questions. When monitoring, she can ask herself, "Did I correctly answer the question?". When controlling, she can, for instance, reread text sections to improve understanding or try to solve a math question for which she is unsure once more before submitting her response. If Anna is skilled at monitoring and controlling her text comprehension but struggles to identify errors in math assignments, her comprehension skills will likely improve, but her math performance may suffer (De Bruin & Van Gog, 2012).

Development of procedural metacognition

Procedural metacognition develops as early as during the preschool years (see Godfrey et al., 2023; Roebers, 2017; and Schneider et al., 2022 for reviews). During elementary school, monitoring skills undergo further development (Godfrey et al., 2023; Roebers, 2017). For example, Van Loon et al. (2017) showed that between the ages of eight and 11, children's ability to recognize their incorrect test responses improved, such that children became less overconfident for errors.

Moreover, children's control skills improve from mid-elementary school onwards. Approximately from the age of eight, children can strategically choose which items they need to restudy to optimize learning (De Bruin et al., 2011; Dufresne & Kobasigawa, 1989; Metcalfe & Finn, 2013; Van Loon et al., 2017). Further, skills to strategically withdraw or submit test responses develop during mid-elementary school. Approximately at the age of eight, children decide to withdraw more incorrect than correct answers, and the accuracy of maintenance/withdrawal decisions improves between the ages of eight and 11 years (Bayard et al., 2021; Roebers et al., 2009).

Most studies on children's metacognition are cross-sectional and only use one task, for instance, picture learning (Liu et al., 2018), paired associates (Metcalfe & Finn, 2013), concept learning (Van Loon et al., 2017), or comprehension of videos or expository texts (De Bruin et al., 2011; Roebers et al., 2009). These studies mainly aimed to describe developmental improvements for different monitoring and control aspects. Although well-suited for estimating general developmental tendencies over time, such approaches might conceal differential developmental patterns of monitoring and control skills.

Heterogeneity of metacognition

In the current study, by combining metacognition measures, we aim to understand how children differ when working on learning tasks (i.e., a paired-associate language learning task and a text-comprehension task) requiring continual monitoring and control. Monitoring and control processes, despite being closely intertwined, appear conceptually distinct. Monitoring involves judging the alignment between knowledge and task goals, while control entails making choices for optimizing performance. Studies show both relations and dissociations between them (e.g., Grainger et al., 2016; Peng & Tullis, 2021; Qui et al., 2018), and from a developmental perspective, monitoring skills appear to precede control skills in childhood (Destan et al., 2014).

Within samples, children of homogenous age groups exhibit substantial heterogeneity in monitoring accuracy and restudy decisions. Follow-up data analyses by Van Loon & Oeri (2023) revealed a moderate to strong overall relation (-0.49 gamma correlation) between fourth graders' CJs and restudy decisions. However, children appeared heterogeneous in how monitoring was translated into control: 50.6% restudied < 20% of items judged as not learned, while 21.5% restudied > 80%. Individual differences in motivation, task value (Wigfield, 1994), and learning goals appear to influence restudy choices (Ariel et al., 2009). Further, children appear to vary in their decisions to maintain or withdraw test responses. Follow-up data analyses by Van Loon et al. (2024) showed that overall, fourth graders maintained 74.5% correct and 32.7% of their incorrect answers when learning texts. However, decisions about incorrect answers varied greatly: 30.9% submitted < 20% of the incorrect responses, while 25.7% submitted ≥ 50%, despite awareness of point deductions when submitting errors.

Studies using multiple tasks to obtain insights into the generalizability of metacognitive skills show that, for adults, monitoring for memory tasks has generalizable elements (Lehmann et al., 2022; Mazancieux et al., 2020). Also, in middle childhood, metacognition appears to some extent to be generalizable (Kleitman & Moscrop, 2010; Van de Stel & Veenman, 2010). For instance, when eight-to-nine-year-old children completed arithmetic and spelling tasks, Bellon et al. (2020) found that monitoring accuracy measures were moderately correlated across tasks. However, particularly for young elementary school children, there may be heterogeneity in the extent to which metacognitive skills are generalizable. Between the ages of eight and ten, metacognitive skills appear to transition from being rather task- or domain-specific to more generalizable across tasks and domains (Bellon et al., 2020; Geurten et al., 2018). In sum, differences between children are likely to exist in (a) their accuracy and effectiveness of monitoring and control, (b) the interplay between monitoring and control, and (c) the generalizability of metacognitive skills across tasks.

In this study, latent profile analysis (LPA) is used to simultaneously consider multiple measures of procedural metacognition to understand how metacognitive monitoring and control processes co-occur within children. Focusing on metacognition profiles rather than mean values for single variables allows us to investigate heterogeneity within individual children's procedural metacognition. For example, some children may struggle with monitoring and control of reading comprehension but not with a paired-associate memory task, or some may be skilled at monitoring but not at controlling. A better understanding of such heterogeneity within children's metacognitive skills may improve the identification of children who are at risk for low metacognition and benefit the design of personalized metacognition support.

A further aspect to consider when aiming to understand heterogeneity in procedural metacognition is the extent to which individual development patterns differ between children over time. The few longitudinal studies on children's development of metacognition mainly focus on metacognitive skills improvement for one specific task as an effect of age (e.g., Fandakova et al., 2017; Roebers & Spiess, 2017). This research, however, does not address to what extent children identified as having proficient metacognition maintain high levels over time and to what extent children who initially lag behind persist in having lower metacognitive skills.

The novelty of the present approach lies in investigating children's membership in metacognitive profiles over time. Building on longitudinal research into the stability and change of self-evaluations such as self-esteem and self-concept (e.g., Putnick et al., 2020; Trzesniewski et al., 2003), we employed an annual measurement interval. This allowed us to explore the extent to which children's metacognitive skills remain stable or change over a year. By investigating the changes in children's profile classifications over time, we can obtain insights into whether metacognitive deficits are temporary and resolved within a year or represent a persistent developmental lag where children with initial metacognitive deficits remain classified as having low procedural metacognitive skills.

Present research

The present study includes multiple metacognition indicators (i.e., confidence judgments, restudy selections, and response maintenance/withdrawal decisions). Two age groups (eight-year-old second graders and ten-year-old fourth graders) participated. With our measures of metacognition, we focus specifically on (a) children's ability to recognize errors and assign lower confidence judgments (CJs) to incorrect responses compared to correct ones; (b) children's skills to correct errors by strategically selecting these task items for restudy, and (c) their ability to strategically maintain correct test responses and withdraw their errors (thereby preventing losing points). In the age range under investigation, metacognitive monitoring and control skills develop extensively. There are pronounced changes in children's capacity to recognize and correct errors, to utilize monitoring mechanisms to guide their control decisions and actions, and metacognition transitions from task-specific to a more generalizable skill (Bellon et al., 2020; Geurten et al., 2018; Selmeczy & Ghetti, 2019; Van Loon et al., 2013).

Two verbal learning tasks, text comprehension and Kanji learning, were completed at two measurement points, one year apart. Despite similarities in task phases—learning, testing, monitoring, and controlling—task formats differed. For the text comprehension task, children read short texts at their own pace, with open-ended questions in the test phase. The Kanji task involved paired-associated visual stimuli with fixed learning times and a multiple-choice recognition test. Procedural metacognition indicators—monitoring accuracy, effective restudy selections, and response maintenance accuracy—were calculated per task and measurement point.

Using person-oriented LPAs, we aimed to identify profiles of procedural metacognition. We expected quantitative differences between children in metacognitive skills, including at least one high and one low metacognitive skills profile in both age groups (H1). To foreshadow, because differences appeared between the two grade levels in profile classifications, LPAs are reported separately for each grade. To better understand potential reasons for differences between children's profile memberships, we investigate whether task performance measures were related to metacognition profile classifications. The relationship between metacognition and performance is subject to ongoing debate, with some research arguing for a close connection (showing that higher task performance often correlates with enhanced metacognition, e.g., Roebers & Spiess, 2017), whereas other research suggests that these are dissociable (Howard et al., 2010; Scott et al., 2014) and should be separable in measurement to interpret "purer" measures of metacognition (Fleming & Lau, 2014). For explorative purposes, we relate task performance to the metacognition profile classifications to obtain insights into whether children's retrospective monitoring and control skills (i.e., monitoring judgments and control decisions were made after answering test questions) may be driven by test performance.

By combining two tasks, we can obtain insights into the generalizability of metacognition. This appears to be developing in the age range under investigation; children may differ in the extent to which their levels of monitoring and control are generalizable. Bellon et al. (2020) found evidence for generalizable metacognitive skills between ages eight and 10. We, therefore, expect that we find, particularly for the older age group, that at least some of the identified profiles would show generalizable metacognitive skills (i.e., similar levels of metacognitive ability for the text task and the Kanji task, respectively, H2).

Longitudinal measurements (one year apart) provide insights into the developmental trajectories of metacognition profiles and the probabilities of changing profile membership. We aim to understand if identified profiles remain similar or change over time (potentially due to biological maturation, repeated task and testing experiences, and schooling experiences). Further, we examine to what extent children's profile memberships develop and change over time. This way, we can obtain insights into how often children with a high level of metacognition remain high, and children initially classified with deficits in monitoring and/or control catch up over time or continue to exhibit metacognitive deficits. Based on findings showing moderate relations between children's level of confidence across two time points (Roebers & Spiess, 2017), we hypothesize that children classified with accurate metacognition tend to remain high in metacognition, and children poorer in metacognition also tend to stay in lower-metacognition profiles over time (H3).

Methods

Participants

At the first measurement point (M1), 327 children participated: 151 second-graders (*M* age 7.61 years, *SD* 0.50, 50% female) and 176 fourth-graders (*M* 9.62 years, *SD* 0.49, 46.38% female). They attended primary schools in the German-speaking part of Switzerland. Of these participants, 325 completed the Kanji task, and 304 completed the text task. At M2, which took place one year later, 308 children participated: 138 second-graders (then in third grade) and 170 fourth-graders (then in fifth grade). Of these participants at M2, 307 children completed the text task and 303 the Kanji task; 279 children completed both the text and the Kanji tasks at both measurement points.

All children were used to follow school instructions in the German language; 76.47% had German as a native language, of the children with other native languages, Albanian (7.19%), Italian (1.96%), and Hungarian (1.96%) were most common. At M2, the children had changed to the next school year and were thus in third and fifth grade. However, for clarity, we will refer to these profiles as second and fourth-graders throughout the paper.

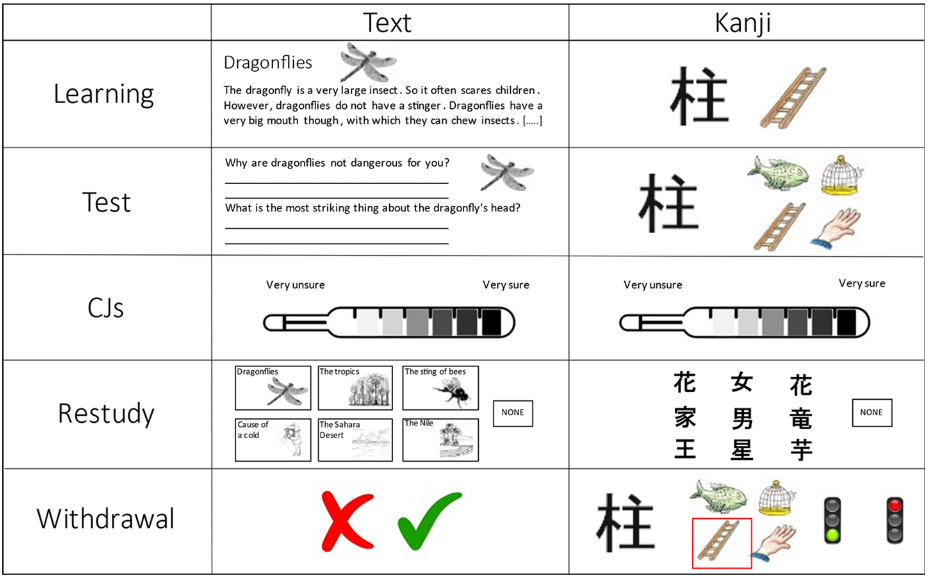
Note that children participated in a large project including seven measurements in one year: In Fall Year 1 with a Kanji and text task, in Spring Year 1 with a Kanji, text, and secret code task, and in Fall Year 2 with a Kanji and text task. Due to the aims of this study, the measurements in the Fall of Year 1 (M1) and one year later in the Fall of Year 2 (M2) were used. Findings about the overarching project are reported in detail for the Kanji tasks by Roebers et al. (2019), and the text task is reported by Steiner et al. (2020). The current approach is the first to use a latent profile approach to combine findings for the text and Kanji tasks cross-sectionally and longitudinally.

Materials and procedure

Children were tested in their school classes. At each measurement point (one year apart), they completed the text and Kanji tasks during two sessions with approximately one week in between. The order of completion was counterbalanced per class. Both tasks consisted of a learning, test, monitoring (CJs), and control phase (restudy selections and responses maintenance/withdrawal). Figure 1 shows these phases. All materials were previously piloted with a different sample of children.

Text comprehension task

Six expository texts were read in randomized order on a tablet; text reading was self-paced. The texts had the same topics but were longer and more difficult for the fourth than the second graders (average word count 80.9, *SD* = 6.2 for second and 127.4, *SD* = 7.0 for fourth grade). More details about the texts are reported by Steiner et al. (2020). After reading, the children received a booklet with test questions. They answered twelve open-ended questions (two per text). For each text, a comprehension question required a sentence as an answer, and a detail question required a single-word response. After giving each answer, they monitored how certain they were that their answer was correct by making a CJ on the 7-point thermometer (adapted from Koriat & Shitzer-Reichert, 2002) printed right of each question. When introducing the thermometer at the first testing session, we asked, “What color is grass?” and prompted them to indicate their confidence level. Children unanimously selected the highest point on the scale (red, very sure) for this known answer. When then asked, “How many hairs do I have on my head?” (for which they made a guess), they uniformly chose the blue side (unsure). These instructions ensured that children all received the same instructions about how to use the scale, making our measures of monitoring accuracy comparable across participants. After making confidence judgments, children could select



The dragonfly is a very large insect. So it often scares children. However, dragonflies do not have a stinger. Dragonflies have a very big mouth though, with which they can chew insects. [.....]

Why are dragonflies not dangerous for you?

What is the most striking thing about the dragonfly's head?

Fig. 1 Procedure

texts for restudy; the text topics were presented on a page in a 3×2 grid format (as recommended by Thiede & Dunlosky, 1999). Children could put a checkmark next to the text they would like to restudy (note that they did not restudy these texts). In the end, children could decide which responses they wanted to maintain and withdraw. They were informed that a+1, -3 bonus-to-penalty ratio (based on Roebers et al., 2009) would be used to award or deduct points, such that they would earn one point for maintaining correct answers, and three points would be deducted for maintaining incorrect answers. They then looked back at their test answers and decided for each of their responses whether they liked to maintain (by adding a check mark) or withdraw them (by crossing the answer out).

Kanji learning task

Children learned 12 (second graders) or 16 (fourth graders) Japanese characters (Kanjis, as used by Destan et al., 2014) and their meanings presented as pictures. The Kanji materials are reported in more detail by Roebers et al. (2019). The task was completed on a tablet computer. When learning, each Kanji and its meaning were randomly shown for 5 s. For the recognition test, one Kanji at a time was presented together with four alternatives (the correct alternative and three randomly selected alternatives that had appeared in the learning phase). Children had to select the image depicting the meaning of the Kanji. There was no time limit to complete the test. After choosing an answer, a red frame surrounded it, and the children made a monitoring judgment (CJ). They answered, “How sure are you that you have chosen the correct picture?” by clicking on a 7-point scale ranging from very unsure to very sure, illustrated with the colored thermometer scale (adapted from Koriat & Shitzer-Reichert, 2002; the thermometer was introduced similarly to the text task.) Then, participants could select which Kanjis they would like to restudy. The Kanjis were presented in a grid format (a 4×3 grid for the second graders and a 4×4 grid for the fourth graders); they could click on the Kanji they liked to restudy and could also select a "none" button.

In the end, participants were instructed that they could decide to maintain or withdraw their answers. They were informed that the + 1, -3 bonus-to-penalty ratio would be used to award or deduct points. Answers appeared one at a time. They could maintain or withdraw their answers by selecting a green or red traffic light.

Analyses

For each task to assess performance, the percentage of correct test responses was calculated. To assess monitoring accuracy, discrimination was calculated as the difference in the mean of confidence for accurate and inaccurate answers (Roebers, 2002). Discrimination could range between −6 (very inaccurate discrimination, i.e., consistently being highly certain for incorrect responses and very uncertain for correct responses) to + 6 (perfectly accurate discrimination, consistently being highly certain for correct responses and very uncertain for incorrect responses.[1] To assess the effectiveness of restudy selections, for the Kanji task, the percentage of restudy selections for items with incorrect responses was calculated per child. For the text task, the percentage of restudy selections for texts for which at least one of the two questions was answered incorrectly was calculated. That is, restudy selections were effective when many of the not-well-learned task materials were selected for further study. Moreover, to investigate decision accuracy, the percentage of correct response maintenance/withdrawal decisions (accurate answers maintained and inaccurate responses withdrawn) was calculated for each child. Monitoring accuracy, effective restudy, and maintenance/withdrawal decision accuracy measures were converted into z-scores for all analyses. For full sample analyses, z-scores were calculated for the whole sample; for the reported follow-up analyses per grade level, z-scores were calculated separately per age group.

Latent profile analyses were conducted with LatentGold software (Vermunt & Magidson, 2005, 2021). Only children with complete data sets per measurement point (i.e., who completed both the text and the Kanji task) were included in these analyses. To test whether we would find different profiles of metacognition (addressing H1), we performed time-specific LPAs for both measurement points separately, using a stepwise procedure from a one to six-profile solution. The LPAs included the z-scores for the constructs: (a) monitoring accuracy texts, (b) effective restudy selections for not well-understood texts, (c) decision accuracy of response maintenance/withdrawal text task, (d) monitoring accuracy Kanji, (e) effective restudy Kanji, and (f) decision accuracy of response maintenance/withdrawal Kanji. For the LPAs, the statistics and p-values for the VLMR (Vuong-Lo-Mendell-Rubin Likelihood

[1] We calculated three measures of monitoring accuracy: discrimination (Roebers, 2002), gamma (Nelson, 1984), and individual beta-values extracted from a mixed effects model using the approach suggested by Murayama et al. (2014). These measures were all strongly correlated with each other, indicating that they capture nuances of one underlying construct. Specifically, the correlations of the discrimination measure with the gamma measure ranged between 0.58 (text task at M2) and 0.77 (Kanji task at M1); the correlations between the discrimination measure and the beta measure ranged between 0.63 (text task at M2) and 0.75 (Kanji task at M2). Further, the measures of discrimination did not appear to be confounded with task performance; correlations between monitoring and performance were low, ranging between −0.06 (Text task M2) and 0.21 (Kanji task M1). We therefore considered the discrimination measure suitable for our research purposes. In this manuscript, we do not further report and interpret the beta and gamma measures. However, for transparency, the gamma and beta values are included in a supplementary data file available on OSF, link https://osf.io/nx2wp/wiki/home/?view_only=9623bc6ce8a24b91ad53c10333f1305e.

Ratio Test) are reported to inform whether adding more profiles significantly improves the model fit. To select the best fitting profile solution, the solutions of the LPAs were compared, and the most parsimonious fitting model was selected, which displayed the smallest BIC values and an entropy value higher than 0.70 (as recommended by Nylund et al., 2007). Further, when conducting LPAs with relatively small sample sizes (lower than 300), a larger number of profiles can result from over-extraction, and small profiles may be challenging to interpret (Nylund-Gibson & Young Choi, 2018). Therefore, we only selected profile solutions for which the smallest profile had a prevalence of at least 10% of the sample. Note that for our initial LPAs, data for second- and fourth-graders was combined, and to account for potential differences between grades, grade level (2 or 4) was included as a nominal, active covariate. This way, we could benefit from the larger sample size (combining both grade levels) and still investigate whether profile classifications would differ for both age groups. To foreshadow, the covariate grade level appeared significant, indicating that profile classifications differed for both age groups. Therefore, follow-up LPAs were conducted for both grade levels separately.

For the LPA profile solutions for both grade levels, we addressed to what extent the identified metacognition profiles differed in task performance for the text and the Kanji task using a three-step approach (with a BCH adjustment). This approach preserves the LPAs' classification uncertainty when relating profile membership to the distal task performance variables (Bakk & Kuha, 2021; Nylund-Gibson et al., 2019).

The data for the children with complete data sets (i.e., including the two measurement points both for the Text task and Kanji tasks) were used for the analyses investigating longitudinal profile changes (for the second graders, 95 children had complete longitudinal data sets; for the fourth graders, 123 children had complete longitudinal data sets). As profile classifications differed for both grades, we separately calculated the within-person profile change probabilities across time points for both age groups. This enabled us to investigate to what extent children initially classified with higher levels of metacognition would remain higher in metacognition (relative to their peers) over time and vice versa (testing H3). With Chi^2 tests, we investigated whether the pattern of changes between profiles from M1 to M2 could be considered random or if this pattern significantly deviated from what would be expected by chance. That is, with these tests, we address whether being classified into a certain profile at M1 would be accompanied by a higher/lower likelihood than chance of being classified into a specific profile one year later at M2). Significant Chi^2 results would indicate that certain movement patterns were more or less probable than expected by chance alone. For significant omnibus Chi^2 tests, the standardized residuals were inspected for each profile change type to follow up on which specific movements between profiles were more or less likely to happen than would be expected by chance alone.

The full dataset and supplementary materials (including tables showing correlations between variables and results of the LPAs for the total sample) are available online: https://osf.io/nx2wp/?view_only=9623bc6ce8a24b91ad53c10333f1305e.

Results

Before testing our hypotheses, descriptive statistics for performance, monitoring, restudy, and response maintenance/withdrawal are presented. Preliminary analyses then investigate differences in means between age groups and differences in means between measurement points. The online supplementary materials show tables with correlations between the LPA-indicator variables and task performance for M1 and M2.

Descriptive statistics and preliminary analyses

For both age groups, Table 1 shows the mean values for performance, CJs for correct and incorrect responses, monitoring accuracy (i.e., CJ discrimination), the percentage of restudy selections, the percentage of effective restudy selections for not-well-learned task materials, the percentage of maintained responses, and decision accuracy (indicated by the percentage of accurate maintenance/withdrawal decisions), for both tasks (text and Kanji) and both measurement points. This table indicates significant differences between age groups for both measurement points.

Table 1 Mean performance, CJs, restudy selections, and response maintenance/withdrawal decisions

Measured Variables	Measurement Point 1		Measurement Point 2	
	Second Graders	Fourth Graders	Second Graders	Fourth Graders
Text Task				
Mean Performance (% correct)	55.16 (23.83)	57.48 (23.96)	59.80 (19.20)**	45.14 (22.02)**
Mean CJs correct responses	5.44 (1.31)	5.24 (1.25)	5.52 (1.06)	5.38 (1.32)
Mean CJs incorrect responses	4.68 (1.59)*	4.06 (1.37)*	3.44 (1.45)	3.30 (1.32)
Discrimination CJs	0.81 (1.40)	1.10 (1.25)	2.06 (1.34)	2.12 (1.43)
Restudy Selections (%)	28.64 (28.60)**	41.26 (28.52)**	32.73 (31.63)*	41.87 (32.57)*
Effective Restudy Selections (%)	27.85 (35.95)*	36.24 (36.84)*	40.64 (38.25)	43.50 (37.93)
Maintained responses (%)	72.57 (24.72)**	60.49 (23.34)**	61.22 (18.24)*	50.60 (19.80)*
Decision Accuracy (%)	60.66 (22.54)	64.96 (18.10)	72.36 (19.06)	71.49 (17.78)
Kanji Task				
Mean Performance (% correct)	44.64 (17.96)**	54.69 (16.22)**	61.19 (20.36)*	68.04 (20.30)*
Mean CJs correct responses	5.01 (1.41)*	5.33 (0.96)*	5.49 (1.06)	5.70 (1.11)
Mean CJs incorrect responses	4.06 (1.51)	4.09 (1.22)	3.99 (1.48)	4.15 (1.58)
Discrimination CJs	0.95 (1.00)*	1.47 (1.24)*	1.25 (0.66)	1.56 (1.16)
Restudy Selections (%)	40.04 (35.90)	46.70 (31.21)	45.49 (35.97)*	35.57 (31.64)*
Effective Restudy Selections (%)	43.99 (41.28)*	55.95 (42.67)*	57.99 (28.04)	49.96 (38.36)
Maintained responses (%)	55.53 (25.25)	52.68 (22.00)	56.32 (22.32)	59.16 (21.42)
Decision Accuracy (%)	55.75 (17.60)**	65.74 (14.56)**	68.32 (14.97)*	72.71 (15.32)*

Note Mean values for performance, CJs for correct and incorrect responses, CJ discrimination, restudy selections (%), effective restudy (%), maintained responses (overall %), and decision accuracy indicating the percentage of maintained correct responses and withdrawn incorrect responses. Mean values are presented for both age groups (second and fourth graders), tasks (text and Kanji), and measurement points — SDs of the mean in parentheses. Significant differences between age groups are indicated with * when $p<.05$ and ** when $p<.001$

As shown in Table 1, for the text task at M1, *t*-tests showed significant differences between age groups in CJs for incorrect responses (second > fourth graders), the number of restudy selections (second < fourth graders), effective restudy selections (second < fourth graders), and the number of maintained responses (second > fourth graders). For the Kanji task at M1, *t*-tests showed significant differences between age groups in mean performance (second < fourth graders), CJs for correct responses (second < fourth graders), CJ discrimination (second < fourth graders), effective restudy selections (second < fourth graders) and decision accuracy (second < fourth graders).

Further, Table 1 shows that at M2, there were significant differences between age groups for the text task in mean performance (second < fourth graders), the number of restudy selections (second < fourth graders), and the number of maintained responses (second < fourth graders). For the Kanji task, there were significant differences between age groups in mean performance (second < fourth graders), the percentage of restudy selections (second > fourth graders), and decision accuracy (second < fourth graders).

Moreover, with paired *t*-tests was investigated whether, for both age groups, the means differed across measurement points. For the text task, significant differences were found for second graders for performance (M1 < M2, $p = .01$), CJs for incorrect responses (M1 > M2, $p < .001$), CJ discrimination (M1 < M2, $p < .001$), effective restudy selections (M1 < M2, $p = .04$), response maintenance (M1 > M2, $p < .001$) and decision accuracy (M1 < M2, $p < .001$). For fourth graders, for the text task, there were differences between measurement points in performance (M1 > M2, $p < .001$), CJs for incorrect responses (M1 > M2, $p < .001$), CJ discrimination (M1 < M2), response maintenance (M1 > M2, $p < .001$), and decision accuracy (M1 < M2, $p = .002$).

For the Kanji task, comparisons between measurement points for second graders showed significant differences in mean performance (M1 < M2, $p < .001$), CJs for correct responses (M1 < M2, $p = .001$), CJ discrimination (M1 < M2, $p < .001$), restudy for incorrect responses (M1 < M2, $p = .003$), and decision accuracy (M1 < M2, $p < .001$). For the fourth graders, for the Kanji task, there were significant differences between measurement points in task performance (M1 < M2, $p < .001$), CJs for correct responses (M1 < M2, $p < .001$), CJ discrimination (M1 < M2, $p = .003$), the overall number of restudy selections (M1 > M2, $p < .001$), the number of restudy selections for incorrect responses (M1 > M2, $p < .006$), the number of maintained responses (M1 < M2, $p < .001$) and decision accuracy (M1 < M2, $p < .001$).

Profiles of metacognition at measurement 1

The supplementary materials show the model fit indices for the full sample for the latent profile solutions with 1–6 latent profile profiles. Although the BIC value was lowest for the 6-profile solution, the smallest profile had a prevalence of 4.9% of the sample. The 5-profile solution appeared to have a high entropy, and all identified profiles had prevalences of at least 16.4%. Therefore, this 5-profile solution was selected. Profile prevalences and mean values for the metacognition measures per profile are shown in the supplementary materials. The Wald test showed a significant effect of the covariate grade level, $\chi^2(1) = 17.37, p = .002$, indicating that classification outcomes differed for both age groups (grade differences per profile are described in the supplementary materials). Therefore, LPAs were subsequently calculated for both grade levels separately.

Profiles for second graders

Table 2 shows the fit indices for latent profile solutions with 1–6 profiles for the second graders. The BIC value was the smallest for the 3-profile solution, entropy was high, and the 3-profile solution fitted better than the 2-profile solution. Therefore, this solution was selected. The mean values for monitoring accuracy, restudy effectiveness, and decision accuracy are presented in Table 3. Further, Fig. 2 shows the z-scores for metacognition for both tasks for the three profiles. The largest profile (54.3% of the included 2nd graders, $n=63$) had overall low to average metacognition; particularly for texts, metacognitive skills were low. This profile is called the *variable metacognition* profile. Profile 2 (35.3%, $n=41$) had high metacognition for the text task; metacognition was average for the Kanji task. This profile is referred to as the *high text metacognition* profile. Profile 3 was a small profile (10.3%, $n=12$); children in this profile had low metacognition for the text task, particularly

Table 2 Model fit indices from latent profile analyses for both grade levels for the solutions with a maximum of 6 latent profile profiles for measurement point 1 and measurement point 2

Number of profiles	Log L	BIC	AIC	VLMR	*p*-value VLMR	Entropy R^2
Measurement Point 1						
Second Graders						
1	-964.03	1985.11	1952.06			1
2	-835.09	1789.02	1720.18	257.88	0.000	0.95
3	-803.44	1787.52	1682.88	63.30	0.000	0.93
4	-773.41	1789.25	1648.82	60.07	0.003	0.91
5	-743.17	1790.57	1614.34	60.48	0.001	0.95
6	-712.85	1791.73	1579.71	60.63	0.000	0.95
Fourth Graders						
1	-1220.23	2500.26	2464.45			1
2	-1120.15	2364.88	2290.29	200.16	0.000	0.95
3	-1040.82	2271.02	2157.64	158.65	0.025	0.92
4	-1009.66	2273.48	2121.31	62.33	0.000	0.88
5	-980.00	2278.96	2088.00	59.30	0.000	0.91
6	-961.95	2307.65	2077.91	36.10	0.065	0.91
Measurement Point 2						
Second Graders						
1	-1046.97	2151.79	2117.95			1.00
2	-936.94	1994.38	1923.88	220.08	0.000	0.98
3	-836.95	1857.08	1749.91	199.97	0.000	0.96
4	-806.56	1858.95	1715.11	46.97	0.000	0.92
5	-779.30	1867.10	1686.60	72.99	0.000	0.95
6	-758.47	1888.11	1670.95	37.25	0.010	0.95
Fourth Graders						
1	-1186.07	2431.77	2396.13			1.00
2	-1111.66	2347.57	2273.32	148.81	0.000	0.95
3	-1084.07	2357.00	2244.15	105.92	0.000	0.82
4	-1027.06	2307.58	2156.12	70.41	0.000	0.92
5	-1006.17	2330.42	2140.35	55.53	0.001	0.90
6	-987.20	2357.08	2128.40	17.31	0.154	0.91

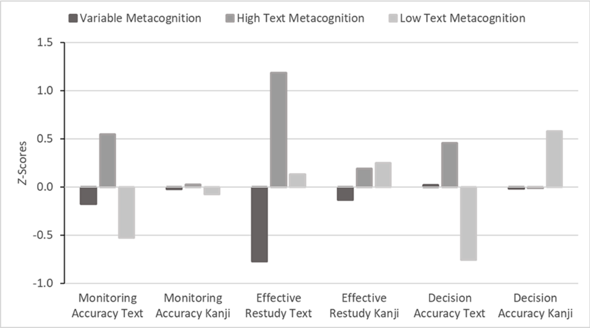
Note Log L=Log Likelihood; AIC=Akaike Information Criteria; BIC=Bayesian Information Criteria, VLMR=Vuong-Lo-Mendell-Rubin Likelihood Ratio Test

Table 3 Mean scores (unstandardized) for monitoring accuracy, effective restudy, and decision accuracy for the identified profiles for both grades at both measurement points

Profile	Monitoring Accuracy Text	Effective Restudy Text	Decision Accuracy Text	Monitoring Accuracy Kanji	Effective Restudy Kanji	Decision Accuracy Kanji
Measurement Point 1						
Second Graders						
Variable Metacognition	0.56 (1.44)	0.00 (0.00)	0.61 (0.20)	0.93 (0.83)	0.38 (0.41)	0.55 (0.19)
High Text Metacognition	1.58 (1.33)	0.70 (0.26)	0.71 (0.13)	0.97 (1.12)	0.52 (0.42)	0.56 (0.14)
Low Text Metacognition	0.06 (0.47)	0.33 (0.14)	0.44 (0.10)	0.88 (1.29)	0.54 (0.40)	0.66 (0.13)
Fourth Graders						
Low Kanji Monitoring	1.14 (1.19)	0.57 (0.26)	0.65 (0.13)	0.98 (0.86)	0.43 (0.28)	0.65 (0.15)
At-Risk Metacognition	0.74 (1.01)	0.00 (0.00)	0.60 (0.19)	1.21 (0.96)	0.43 (0.37)	0.64 (0.15)
Proficient Metacognition	1.71 (1.45)	0.73 (0.30)	0.68 (0.18)	1.61 (0.89)	1.00 (0.00)	0.67 (0.13)
Measurement Point 2						
Second Graders						
Effective Restudy	2.05 (1.39)	0.54 (0.39)	0.72 (0.17)	1.73 (1.22)	1.00 (0.02)	0.69 (0.14)
Low Kanji Monitoring	2.15 (1.29)	0.43 (0.29)	0.77 (0.17)	1.17 (1.11)	0.41 (0.25)	0.64 (0.13)
Ineffective Restudy	1.99 (1.29)	0.00 (0.00)	0.65 (0.21)	1.58 (1.41)	0.01 (0.03)	0.71 (0.17)
Fourth Graders						
Average Metacognition	2.39 (1.36)	0.58 (0.25)	0.72 (0.15)	1.63 (1.24)	0.49 (0.36)	0.73 (0.14)
Ineffective Restudy	1.75 (1.41)	0.00 (0.00)	0.73 (0.17)	1.51 (1.09)	0.38 (0.37)	0.73 (0.12)
Effective Restudy	1.81 (1.52)	1.00 (0.00)	0.68 (0.18)	1.43 (1.10)	0.99 (0.04)	0.69 (0.13)

Note Standard deviations in parentheses

Fig. 2 Latent profile solutions for the first measurement point for second graders



for monitoring and decision accuracy, and average to high metacognition for the Kanji task. This profile is referred to as the *low text metacognition* profile. Profiles differed in performance on the text task, Wald test $\chi^2(2) = 9.46$, $p = .008$. Post hoc comparisons showed that the low text metacognition profile had lower performance on the text task than the variable metacognition profile ($p = .004$) and the high text metacognition profile ($p = .005$); there was no difference between high text metacognition and variable metacognition ($p = .79$). There were no differences between profiles in performance on the Kanji task ($p = .14$).

Metacognitive skills did not appear generalizable across tasks for second graders. Profiles differed most in their ability to monitor and control text comprehension, with one profile showing particularly high skills and a small profile showing particularly low metacognition for text.

Profiles for fourth graders

Table 2 shows the fit indices for the solutions with 1–6 profiles for the fourth graders. The 3-profile solution had the smallest BIC value and high entropy and fitted the data better than the 2-profile solution. Therefore, this solution was selected. Table 3 shows the mean scores for the metacognition measures for the identified profiles; Fig. 3 shows the *z*-standardized means. The largest profile (39% of the sample, $n = 57$) showed average metacognition for the text task; however, monitoring was lowest, and also restudy was low for the Kanji task. This profile is referred to as the *low Kanji monitoring* profile. The second profile (35.6%, $n = 52$) showed consistently low metacognition for both the text and the Kanji task and is referred to as the *at-risk metacognition* profile. The third profile (25.3%, $n = 37$) has high monitoring and control for both tasks and is called the *proficient metacognition* profile.

There were no differences between profiles in performance on the text task ($p = .09$), but for the Kanji task, the profiles differed in performance, $\chi^2(2) = 7.94$, $p = .019$. The low Kanji monitoring and the at-risk metacognition profiles had lower performance on the Kanji task than the proficient metacognition profile (*p*s 0.009 and 0.015, respectively). There was no difference in Kanji task performance between the low Kanji monitoring and the at-risk metacognition profiles ($p = .76$).

For fourth graders, metacognitive skills appeared generalizable across tasks for two of the three identified profiles, i.e., for the at-risk and proficient metacognition profiles. For the low Kanji monitoring profile, for Kanji, monitoring was less accurate and restudy less effective than for the text task.

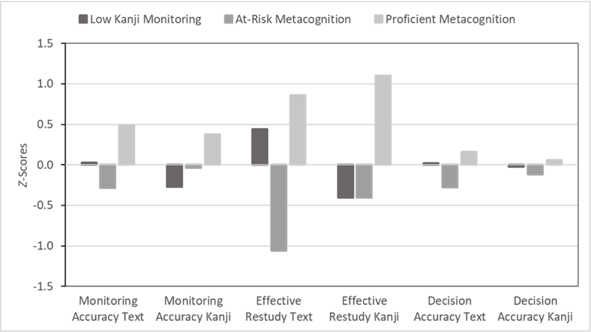


Fig. 3 Latent profile solutions for the first measurement point for fourth graders

Profiles of metacognition at measurement 2

For the LPA for the total sample, the fit indices for the solutions with 1–6 latent profiles are shown in the supplementary materials. Although the 5-profile solution had the smallest BIC value, the smallest profile only contained 7.7% of the sample. The 4-profile solution also had a high entropy value, and all profile prevalences were higher than 10%. Therefore, this solution was selected. The supplementary materials show the profile prevalences and the scores for monitoring accuracy, effective restudy, and decision accuracy for the classified profiles for the 4-profile solution. Because the grade level covariate affected the classification outcomes, $\chi^2(1)=17.20$, $p<.001$ (grade differences are described in the supplementary materials), follow-up LPAs were conducted for both grade levels separately.

Profiles for second graders

Table 2 shows the profile classifications for M2 for the second-graders. BIC values were lowest for the 3-profile solution, entropy was high, and this solution fitted the data better than the 2-profile solution. Further, all profiles had acceptable prevalences. Therefore, the 3-profile solution was selected. Table 3 shows the mean values, and Fig. 4 shows the z-scores for the metacognition measures. The largest profile (40.3%, $n=50$) showed average monitoring and decision accuracy and highly effective restudy for both tasks. This profile is referred to as the *effective restudy* profile. Profile 2 (35.5%, $n=44$) showed average metacognition for the text task but the lowest monitoring and decision-making and low restudy for the Kanji task and is referred to as the *low Kanji monitoring* profile. The third profile (24.2%, $n=30$) had average to low monitoring and decision accuracy and made ineffective restudy selections for both tasks; this profile is referred to as the *ineffective restudy* profile.

There were no differences between profiles in performance on the text task ($p=.34$) and the Kanji task ($p=.27$), indicating that task performance was not driving profile membership. Moreover, as indicated by Fig. 3, for the effective restudy and the ineffective restudy profiles, indicators of metacognition appeared generalizable across tasks. The level of metacognitive skills did not appear generalizable for the low Kanji monitoring profile (i.e., metacognition was consistently lower for Kanji than for texts).

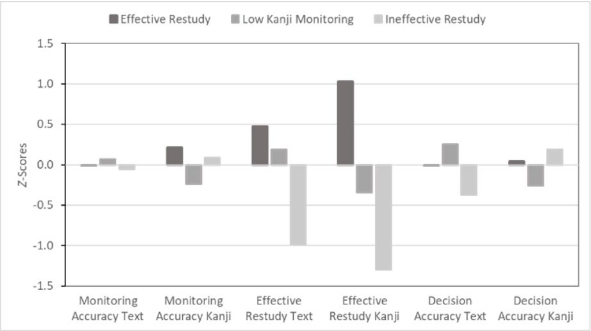


Fig. 4 Latent profile solutions for the second measurement point for second graders

Profiles for fourth graders

For M2, for the fourth graders, as shown in Table 2, the LPAs had the lowest BIC value for the 4-profile solution. However, inspection of the profiles showed that the smallest profile only had a prevalence of 5.6%. The 3-profile solution also had a high entropy, profile prevalences were acceptable (all prevalences > 10%), and the solution fitted better than a 2-profile solution. Therefore, the 3-profile solution was selected.

Table 3 shows the unstandardized scores for monitoring accuracy, effective restudy, and decision accuracy for the three profiles, and Fig. 5 shows the *z*-standardized scores. The first and largest profile (58.3%, $n = 84$) had average monitoring and control skills for the text and the Kanji task. This profile is referred to as the *average metacognition* profile. The second profile (31.3%, $n = 45$) had average-to-low monitoring and decision accuracy for both tasks and restudy was ineffective. This profile is called the *ineffective restudy* profile. The third and smallest profile (10.4%, $n = 15$) had average-to-low monitoring and decision accuracy; however, restudy selections were highly effective. This profile is referred to as the *effective restudy* profile.

Task performance did not affect profile membership; there were no differences between profiles in performance on the text task ($p = .17$) and the Kanji task ($p = .70$). Moreover, for all three profiles, metacognitive skills appeared generalizable across tasks, such that the level of metacognitive skills was average-to-high for both tasks for the average metacognition profile, discrimination was average-to low and restudy particularly low for both tasks for the ineffective restudy profile, whereas restudy was highest for both tasks for the effective restudy profile.

In sum, for both age groups and at both time points, identified profiles showed differences in the quality of metacognitive monitoring and control skills (confirming H1). Additionally, there was heterogeneity in the extent to which metacognitive skills appeared generalizable. There was stronger evidence for H2 (expecting profiles with generalizable metacognitive skills across tasks) at M2 than M1, particularly among older age groups. At M2, all identified profiles showed generalizable metacognitive skills across tasks for fourth graders.

Longitudinal movement patterns between profiles

The extent to which profile membership at M1 was related to profile membership at M2 was inspected for both grade levels separately. The types of changes between profiles over time is shown in Table 4 for second graders and Table 5 for fourth graders. For both grades, the

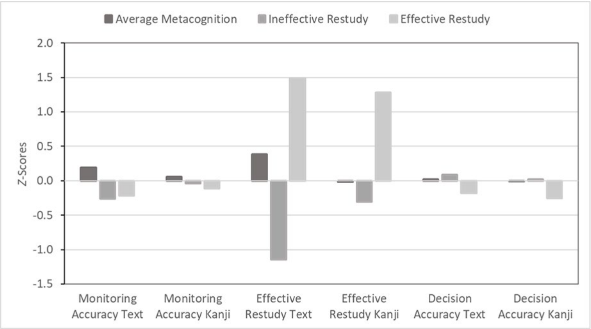


Fig. 5 Latent profile solutions for the second measurement point for fourth graders

Table 4 Probabilities of longitudinal changes between profiles from measurement points 1 to 2 for second graders (in percentage)

Profile Measurement 1	Profile Measurement 2			
	Effective Restudy	Low Kanji Monitoring	Ineffective Restudy	Total *n*
Variable MC	28.8% ($n=15$)	38.5% ($n=20$)	32.7% ($n=17$)	52
High Text MC	51.4% ($n=18$)	37.1% ($n=13$)	11.4% ($n=4$)	35
Low Text MC	75.0% ($n=6$)	12.5% ($n=1$)	12.5% ($n=1$)	8
Total *n*	39	34	22	95

Note MC = Metacognition; *n* indicates the number of observations for each specific longitudinal profile change. The total *n* indicates the number of children with complete data sets (Texts and Kanji tasks completed at both measurement points) who were classified in these profiles

Table 5 Probabilities of longitudinal changes between profiles from measurement points 1 to 2 for fourth graders (in percentage)

Profile Measurement 1	Profile Measurement 2			
	Average MC	Ineffective Restudy	Effective Restudy	Total *n*
Low Kanji Monitoring	76.6% ($n=36$)	19.1% ($n=9$)	4.3% ($n=2$)	47
At-Risk MC	46.5% ($n=20$)	44.2% ($n=19$)	9.3% ($n=4$)	43
Proficient MC	66.7% ($n=22$)	18.2% ($n=6$)	15.2% ($n=5$)	33
Total *n*	78	34	11	123

Note MC = Metacognition; *n* indicates the number of observations for each specific longitudinal profile change. The total n indicates the number of children with complete data sets (Texts and Kanji tasks completed at both measurement points) who were classified in these profiles

changes between profiles from M1 to M2 differed from what would be expected by random chance (second grade, $\chi^2 = 10.90$, $p = 0.028$, fourth grade, $\chi^2 = 12.42$, $p = 0.015$). Follow-up inspections for the second graders did not indicate that specific changes over the course of one year were significantly more or less likely to happen than would be expected by chance. However, for the fourth grade, being classified into the at-risk metacognition profile at M1 was predictive of being classified into the ineffective restudy metacognition profile at M2 (standardized residual = 2.1, $p < .05$). Thus, for fourth-grade children, findings confirm H3. Almost half (44.2%) of the children with low metacognitive skills in fourth grade appeared likely to keep lagging behind and stay at risk for showing metacognitive deficiencies one year later.

Discussion

This study aimed to improve understanding of the development of elementary school children's metacognitive monitoring and control. The novelty of this research is that a latent profile approach was used to combine multiple indicators of metacognition per participant.

This way, metacognition could be investigated as a holistic set of skills rather than focusing on mean values for single variables. Monitoring accuracy (discrimination between CJs for correct and incorrect responses) and control processes (effective restudy selections and the accuracy of response maintenance/withdrawal decisions) were measured in two age groups (second graders and fourth graders) across two tasks (text comprehension and Kanji memory) at two time points one year apart. We investigated whether we could uncover different patterns of metacognition at the individual level (i.e., metacognitive profiles) to understand heterogeneity in children's metacognition, the generalizability of metacognition across two tasks, and heterogeneity in developmental trajectories over time.

As expected (H1), different metacognitive profiles emerged, reflecting heterogeneity in the quality of children's metacognition. For the second graders at M1, three profiles emerged: variable, high text, and low text metacognition profiles. Most second graders appeared to have variable metacognitive skills, and heterogeneity in metacognitive skills was mainly visible for metacognition for the text task. Notably, children with the lowest metacognitive skills for the text task also exhibited the poorest performance on it. Between the ages of 7 and 11, after obtaining text decoding skills, children rapidly develop reading comprehension abilities (Oakhill et al., 2003). The observed differences in metacognition for the text task among second graders possibly mirror individual differences in their developing reading skills.

Also, differences in the level of metacognitive skills were observed for fourth graders, and primary task demands seemed to influence metacognitive skills at M1. Particularly for the Kanji task, fourth-grade children who experienced more difficulties with learning the Kanji stimuli also demonstrated lower metacognitive abilities for this task. Further, in addition to a low Kanji monitoring profile, for fourth graders, an at-risk metacognition profile (with overall low skills) and a proficient metacognition profile (with overall high skills) were identified.

Also, at the second measurement point one year later (M2), three distinct metacognition profiles were identified for second graders (then in third grade). The largest profile was an effective restudy profile. Further, a low Kanji monitoring and an ineffective restudy profile were identified. At M2, for second graders, metacognition seemed more generalizable than at M1, such that the level of metacognition appeared similar across two different tasks for the effective restudy and the ineffective restudy profiles. Moreover, profile classifications were unrelated to task performance. Also, for fourth graders at M2 (who were then in fifth grade), three profiles were found: average metacognition, ineffective restudy, and effective restudy. For all three profiles, metacognitive skills seemed generalizable across the two tasks. Further, task performance did not affect profile classifications for the fourth graders at M2.

Findings thus indicate heterogeneity between children in the quality of metacognition. Moreover, cross-sectional and longitudinal findings indicated development in generalizability of metacognition. For second graders, fewer profiles showed generalizability across tasks than for fourth graders, and for both age groups, at M1, metacognition appeared less generalizable than at M2. Bellon et al. (2020) proposed that the domain generality of metacognition develops around the age of eight. Our findings indicate that when children reached fifth grade and were approximately 11 years old, metacognitive abilities appeared generalizable rather than task-specific for all identified profiles. This indicates that when children cannot monitor and control their text comprehension at this age, they are most likely also unable to monitor and control their Kanji memory.

While profiles show variations in metacognitive skills among children, the identified profile types were not exactly similar for the different grade levels and measurement points. This difference between grade levels and measurement points may indicate that biological maturation and richer educational experiences impact metacognitive development differently as children age. At M1, the metacognition tasks were new, and at M2, the children had experience completing these; the repeated exposure to the metacognition task at M2 could have contributed to observed differences. Further, at M1, profile membership seemed to some extent driven by performance (particularly for the profiles with less generalizable metacognitive skills, low performance on a task was related to low metacognition). However, at M2, the classified profiles appeared to reflect differences in metacognition rather than differences in performance.

The most pronounced difference between the profiles seems to be the extent to which children made effective restudy selections. For children in profiles showing metacognitive deficiencies, most children decided to restudy very few not-learned items; often, they did not restudy anything at all. The high metacognition profiles were characterized by highly effective restudy selections. In the present study, children made restudy selections after answering test questions, and they did not actually restudy to improve their learning. However, in self-regulated learning settings, these highly strategic learning decisions would most likely lead to improved learning scores and the largest long-term learning improvements (Ohtani & Hisasaka, 2018; Van Loon & Oeri, 2023).

This study specifically focused on heterogeneity in children's discrimination between correct responses and errors and the extent to which they decided to correct their errors through restudying and withdrawing these. During the elementary school years, the influence of monitoring judgments on control, and particularly restudy, strengthens (Metcalfe & Finn, 2013). Although our study did not specifically focus on monitoring-based control, it is noteworthy that (as visible in the online supplementary material) children who more accurately monitored their performance also made more effective restudy selections. However, the correlations between monitoring accuracy and effective restudy skills were low to moderate, indicating that while monitoring was related to restudy choices, it was not a perfect relationship. This suggests that other variables, for instance, motivational factors, liking of the study tasks, perceptions about task difficulty, and beliefs about the extent to which they would be able to improve, also affected children's decisions to restudy. Future research could further investigate reasons for heterogeneity in restudy decisions by including motivational variables.

For second graders, overall, the effectiveness of restudy improved for both tasks over one year. However, for fourth graders, mean restudy selections and effective restudy decreased for the Kanji task, and further, performance on the text task appeared to decline over time. These time-related decreases in fourth graders' control and in their learning performance may be explained by declines in intrinsic learning motivation from middle childhood (around age 9) onwards (Gnambs & Hanfstingl, 2015; Gottfried et al., 2001). Interestingly, the LPAs show extensive individual differences among fourth graders in their decisions to restudy, suggesting that only some experienced this motivational decline. Future research should explore how motivational factors impact differences in metacognitive skills among children.

The analyses investigating longitudinal profile changes show how children moved between profiles over one year. These analyses can bring insight into whether children with initially low metacognition at M1 catch up over time or persist with low skills. In second grade, children with lower metacognitive skills tended to catch up over one year, suggesting a pattern of ongoing developmental progression. This indicates that around the age of eight, poor metacognition may indicate delayed development rather than longer-lasting metacognitive deficits. Notably, developmental improvements, especially in restudy effectiveness, were evident over one year for this age group.

However, signs of longer-term challenges in metacognitive development appeared more identifiable in fourth grade. This may indicate that spontaneous metacognition improvement is less likely at this stage. For this age group, deficits in metacognitive skills tended to amplify over time, potentially evolving into developmental risks. These findings seem to imply that, particularly after age 10, children's metacognitive skills should be monitored so that when children appear to have low metacognition, they can be given help earlier rather than later to promote their skills. The children who appeared to be at risk for persistent metacognitive deficits exhibited the most challenges in making effective restudy selections. It thus seems highly unlikely that these children can strategically improve their learning through further study. This may imply that, when aiming to identify whether children have metacognitive challenges, it may be advisable to pay particular attention to children's ability to take appropriate action to learn unknown information and rectify errors.

For fourth graders, 15.4% of the children showed metacognition deficiencies both at M1 (being classified in the at-risk metacognition profile) and M2 (then being classified in the ineffective restudy profile), indicating that they may be at risk for a developmental delay in metacognition. This prevalence seems similar to the prevalence of children in classrooms who have specific learning disorders or disabilities (in the USA 15%, National Center for Education and Statistics, 2024). For these children, most often specific learning disorders (SLD, characterized by difficulties in acquiring reading, writing, and math skills, Morsanyi et al., 2018) and attention deficit hyperactivity disorder (ADHD) are diagnosed. SLD and ADHD seem to be characterized by difficulties in metacognition and effortful control (Butzbach et al., 2021; Tzohar-Rozen et al., 2021). Children diagnosed with a form of SLD (dyscalculia or dyslexia) appear to make less accurate self-monitoring judgments than peers without these disabilities (Baten & Desoete, 2019; Geurten & Lemaire, 2023). Currently, very little research investigates reasons for low procedural metacognition and the potential comorbidity with other learning problems. Future research should prioritize investigating the co-morbidity of low procedural metacognition with ADHD, SLD, and other risk factors for learning difficulties.

One limitation of the present research is the sample size. Although this sample size is larger than in most developmental research, an even larger sample size would have been preferable for LPAs, as it could provide more robust findings and support the interpretation of the small profiles (Spurk et al., 2020). Selecting profile solutions and then naming profiles is a nuanced and somewhat interpretative process. Our chosen profile solutions revealed distinct levels of metacognitive skills, particularly when examining the effectiveness of restudy selections. We selected the 3-factor solutions for reasons of interpretation; although these profile solutions fit the data well, a higher number of profiles may even fit the data better. In this study, we could not further pursue this due to sample size reasons (i.e., very small profiles would then emerge). We took a conservative approach to ensure we would not over-interpret solutions with very small profiles; however, our findings should be evaluated carefully. The 3-factor solutions may be a simplification of reality, and possibly, there is more heterogeneity in metacognition than what appears from our data. The main finding of the present research is that there are heterogeneous groups that clearly differ in the quality of procedural metacognition. The differences between profiles appear most pronounced for children's effectiveness of restudy selections. However, we acknowledge that the reality may be more nuanced, and future research with much larger sample sizes should provide more insights into how many different metacognition profiles would emerge, how this may depend on children's age, how profiles differ from each other, and how profiles may be most adequately named.

When aiming to holistically understand metacognitive skills, a limitation is that we only had verbal metamemory tasks, asking children to memorize materials and then retrieve this information when taking the test. Interpretations about the extent to which the level of metacognitive skills is similar for multiple tasks may be different when verbal tasks are combined with non-verbal tasks (as done, e.g., by Bellon et al., 2020). Moreover, we only had one indicator of monitoring per task (CJs) and two indicators of control (restudy selections and decision accuracy). All these measures of metacognition were retrospective, such that these were collected after children completed test questions. Future research should aim to extend results with more varied task types and other measures of metacognition (e.g., combining prospective and retrospective judgments).

In summary, heterogeneity between children in metacognition is often overlooked. Findings may advance measurement methods to assess procedural metacognition and to distinguish typical from at-risk development. The present study found different profiles of metacognition, varying in the quality of their monitoring and control skills and the generalizability of metacognition across tasks. Children mainly differed in the effectiveness of their restudy selections. Findings contribute to a better understanding of the typical development of metacognitive skills and offer insights for identifying at-risk development. While eight-year-old children with low metacognition tended to catch up over time, ten-year-old children who were low in metacognition were likely to continue falling behind and had limited chances of catching up later. Specifically, for older children, the inability to improve their learning and correct errors through restudy may signal a more prolonged risk. Approximately 15% of the children appeared to be at risk of severely lagging behind in metacognitive development over time. This suggests that educators should be aware of low metacognition in children from age ten onwards. Findings emphasize the importance of early identification of elementary school children with low metacognition and providing individualized support for improvement, given the limited likelihood of spontaneous catch-up.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-024-09400-2.

Acknowledgements The authors like to thank Natalie Bayard, Martina Steiner, and the involved student assistants for their support with task development and data collection.

Funding This research was funded by the Swiss National Science Foundation, Grant Number 100014_165477/1.
Open access funding provided by University of Zurich

Data availability Data and supplementary online materials are available on OSF: https://osf.io/nx2wp/?view_only=9623bc6ce8a24b91ad53c10333f1305e.

Declarations

Ethical approval Ethical approval was obtained from the Faculty of Human Sciences' Ethical Review Board of the University of Bern, Switzerland (Approval No: 2016-08-00004).

Conflicts of interest The authors have no conflicts of interest to report.

References

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432–447. https://doi.org/10.1037/a0015928

Bakk, Z., & Kuha, J. (2021). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 340–362. https://doi.org/10.1111/bmsp.12227

Baten, E., & Desoete, A. (2019). Metacognition and motivation in school-aged children with and without mathematical learning disabilities in Flanders. *Zdm Mathematics Education*, *51*(4), 679–689. https://doi.org/10.1007/s11858-018-01024-6

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

Bellon, E., Fias, W., & De Smedt, B. (2020). Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLOS ONE*, *15*(3), e0229932. https://doi.org/10.1371/journal.pone.0229932

Butzbach, M., Fuermaier, A. B., Aschenbrenner, S., Weisbrod, M., Tucha, L., & Tucha, O. (2021). Metacognition, psychopathology and daily functioning in adult ADHD. *Journal of Clinical and Experimental Neuropsychology*, *43*(4), 384–398. https://doi.org/10.1080/13803395.2021.1940104

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005

De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, *22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

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

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology*, *47*(2), 274–296. https://doi.org/10.1016/0022-0965(89)90033-7

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky, & S. U. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–38). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.14

Fandakova, Y., Selmeczy, D., Leckey, S., Grimm, K. J., Wendelken, C., Bunge, S. A., & Ghetti, S. (2017). Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proceedings of the National Academy of Sciences*, *114*(29), 7582–7587. https://doi.org/10.1073/pnas.1703079114

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

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Hunan Neuroscience*. https://doi.org/10.3389/fnhum.2014.00443. 8.

Geurten, M., & Lemaire, P. (2023). The influence of irrelevant emotionally negative stimuli on early and late retrospective metacognitive judgements. *Quarterly Journal of Experimental Psychology*, 17470218231191516. https://doi.org/10.1177/17470218231191516

Geurten, M., Meulemans, T., & Lemaire, P. (2018). From domain-specific to domain-general? The developmental path of metacognition for strategy selection. *Cognitive Development*, *48*, 62–81. https://doi.org/10.1016/j.cogdev.2018.08.002

Gnambs, T., & Hanfstingl, B. (2015). The decline of academic motivation during adolescence: An accelerated longitudinal cohort analysis on the effect of psychological need satisfaction. *Educational Psychology*, *36*(9), 1691–1705. https://doi.org/10.1080/01443410.2015.1113236

Godfrey, M., Casnar, C., Stolz, E., Ailion, A., Moore, T., & Gioia, G. (2023). A review of procedural and declarative metamemory development across childhood. *Child Neuropsychology*, *29*(2), 183–212. https://doi.org/10.1080/09297049.2022.2055751

Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, *93*(1), 3–13. https://doi.org/10.1037/0022-0663.93.1.3

Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Consciousness and Cognition*, *42*, 65–74. https://doi.org/10.1016/j.concog.2016.03.003

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

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

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

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

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*(3), 325–340. https://doi.org/10.1348/000709910x485719

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

Liu, Y., Su, Y., Xu, G., & Pei, M. (2018). When do you know what you know? The emergence of memory monitoring. *Journal of Experimental Child Psychology*, *166*, 34–48. https://doi.org/10.1016/j.jecp.2017.06.014

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

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

Morsanyi, K., van Bers, B. M., McCormack, T., & McGourty, J. (2018). The prevalence of specific learning disorder in mathematics and comorbidity with other developmental disorders in primary school-age children. *British Journal of Psychology*, *109*(4), 917–940. https://doi.org/10.1111/bjop.12322

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

National Center for Education Statistics (2023). Students With Disabilities. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved 20.3.2024, from https://nces.ed.gov/programs/coe/indicator/cgg

National Center for Education Statistics. (2024). Students with disabilities. U.S. Department of Education, Institute of Education Sciences. Retrieved August 8, 2024, from https://nces.ed.gov/programs/coe/indicator/cgg/students-with-disabilities

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

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

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). The MIT Press.

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440–461. https://doi.org/10.1037/tps0000176

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967–985. https://doi.org/10.1080/10705511.2019.1590146

Nylund, K. L., Asparouhov, T., & Muthen, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*(4), 443–468. https://doi.org/10.1080/01690960344000008

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

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

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

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

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

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

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

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

Roebers, C. M., & Spiess, M. (2017). The development of metacognitive monitoring and control in second graders: A short-term longitudinal study. *Journal of Cognition and Development*, *18*(1), 110–128. https://doi.org/10.1080/15248372.2016.1157079

Schneider, W., Tibken, C., & Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications. In J. J. Lockman (Ed.), *Advances in child development and behavior* (Vol. 63, pp. 273–307). https://doi.org/10.1016/bs.acdb.2022.04.006

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

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

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and how to guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Steiner, M., Van Loon, M. H., Bayard, N. S., & Roebers, C. M. (2020). Development of children's monitoring and control when learning from texts: effects of age and test format. *Metacognition and Learning*, *15*, 3–27. https://doi.org/10.1007/s11409-019-09208-5

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024

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

Tzohar-Rozen, M., Waisel, T., & Shamir, A. (2021). Self-regulated learning skills and the quality of execution of construction tasks: A comparison between young children at risk for specific learning disorders and typically developing peers. *Metacognition and Learning*, *16*, 833–853. https://doi.org/10.1007/s11409-021-09266-8

Van der Stel, M., & Veenman, M. V. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*(3), 220–224. https://doi.org/10.1016/j.lindif.2009.11.005

Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., & Van Merriënboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning*, *8*(2), 173–191. https://doi.org/10.1007/s11409-013-9100-0

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

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

Van Loon, M., Orth, U., & Roebers, C. (2024). The structure of metacognition in middle childhood: Evidence for a unitary metacognition-for-memory factor. *Journal of Experimental Child Psychology*, *241*, 105857. https://doi.org/10.1016/j.jecp.2023.105857

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the Presence of Measurement Non-invariance or Differential Item Functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 356–364. https://doi.org/10.1080/10705511.2020.1818084

Vermunt, J., & Magidson, J. (2005). *Technical Guide for latent GOLD choice 4.0: Basic and Advanced*. Statistical Innovations inc.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*, 49–78. https://doi.org/10.1007/BF02209024