Unraveling the reciprocal effects and micro-development of metacognitive monitoring strategies and monitoring judgments: a micro-genetic diary study

Marion Händel, Ulrike Nett, Donna Bryce, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Händel, Marion, Ulrike Nett, Donna Bryce, and Markus Dresel. 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.

Learning and Individual Differences 122 (2025) 102748

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Unraveling the reciprocal effects and micro-development of metacognitive monitoring strategies and monitoring judgments: A micro-genetic diary study[☆]

Marion Händel[a,*], Ulrike Nett[b], Donna Bryce[b], Markus Dresel[b]

[a] Ansbach University of Applied Sciences, Residenzstr. 8, 91522 Ansbach, Germany
[b] University of Augsburg, Universitätsstr. 10, 86159 Augsburg, Germany

ARTICLE INFO

Keywords:
Metacognitive monitoring strategies
Spontaneous monitoring judgments
Judgment accuracy
Cross-lagged

ABSTRACT

For successful learning, monitoring one's learning behavior and outcomes is essential. While both the use of monitoring strategies and monitoring judgment accuracy are regarded as indicators for metacognitive monitoring, their joint contribution to metacognitive monitoring and their interrelationship is unclear. The current study investigated how monitoring strategies and monitoring judgments (spontaneous as well as prompted judgments) change micro-genetically and influence each other reciprocally during an exam preparation phase. In this diary study, $N = 273$ higher education students completed knowledge tests and surveys on five consecutive days. Results indicated interindividual differences in dependence of prior performance and gender for the accuracy of prompted judgments but not for monitoring strategies or spontaneous judgments. Random intercept cross-lagged panel models revealed that higher quality monitoring strategies led to the spontaneous formation of monitoring judgments, while the quality of monitoring strategies and prompted judgment accuracy did not influence each other significantly.

Educational relevance and implications statement: This study showed that students metacognitively monitor their learning in an intense exam preparation phase via monitoring strategies as well as spontaneous judgments, regardless of prior performance or gender. As their spontaneous judgments were influenced by the quality of their reported monitoring strategies, the training of monitoring strategies seems an ideal starting point to foster metacognitive monitoring in students.

1. Introduction

The study of metacognitive monitoring is pivotal in understanding how individuals assess and regulate their cognitive processes. Metacognitive monitoring is a central component of self-regulated learning (Zimmerman, 1998). According to Zimmerman and Moylan (2009), metacognitive monitoring refers to "informal mental tracking of one's performance processes and outcomes, such as one's learning processes and their effectiveness in producing learning" (p. 303). Only if learners gain up to date and accurate information through monitoring processes, can they regulate their learning accordingly, for example by intensifying or reducing the learning efforts.

Research on metacognitive monitoring typically adopts one of two perspectives: First, in line with models of self-regulated learning (Winne & Hadwin, 2008), the strategy-oriented perspective examines the monitoring of the learning process via strategies. Second, the judgment-oriented perspective emerged from research on metamemory (Nelson, 1990) and focuses on the monitoring of personal performance before, during or after retrieval via judgments. Until recently, those two strands of research were usually studied separately without even considering the other perspective. An integrate model that considers both monitoring strategies as well as monitoring judgments was proposed by Händel and Dresel (2022). Their model suggests that monitoring strategies as well as judgment accuracy contribute to metacognitive monitoring and successful learning and that they are reciprocally related: The use and quality of monitoring strategies should enhance judgment accuracy and

☆ This article is part of a Special issue entitled: 'SRL' published in Learning and Individual Differences.

* Corresponding author.
E-mail addresses: marion.haendel@hs-ansbach.de (M. Händel), ulrike.nett@uni-a.de (U. Nett), donna.bryce@uni-a.de (D. Bryce), markus.dresel@uni-a.de (M. Dresel).

https://doi.org/10.1016/j.lindif.2025.102748
Received 25 September 2024; Received in revised form 18 June 2025; Accepted 24 June 2025
Available online 3 July 2025

vice versa. So far, related research indicates only low correlations of the two types of metacognitive monitoring (Dörrenbächer-Ulrich et al., 2021; Dörrenbächer-Ulrich & Perels, 2023; Händel et al., 2023; Händel & Dresel, 2022). However, previous research is limited to cross-sectional approaches and therefore an important question is, do monitoring strategies and monitoring judgments influence each other during the learning process and if so, how? To address this relationship comprehensively, longitudinal studies are needed as they can better examine the reciprocal influences between monitoring strategies and monitoring judgments over time, providing a clearer understanding of their interplay. Hence, the current study used a (micro-)longitudinal design and considered individual differences in both types of metacognitive monitoring.

1.1. Types and functions of metacognitive monitoring

Metacognitive monitoring is highly relevant for successful learning and has several functions during the knowledge acquisition and retrieval processes. We will illustrate this via an example of a higher education learning situation about the topic of photosynthesis. First, metacognitive monitoring serves to observe and critically reflect upon one's learning behavior, i.e., whether the implemented learning strategies are useful to reach the learning goal (Lan, 2005), in this case, understanding the process of photosynthesis. For example, learners can read a text about photosynthesis and monitor whether highlighting text passages on the impact of the sunlight helps them to understand the relevance of photosynthesis for life on earth. This activity reflects the use of monitoring strategies in the learning phase, with processing as the object of monitoring. Second, metacognitive monitoring serves to identify potential knowledge gaps or difficulties in understanding (Baars, 2023; Kruger & Dunning, 1999). For example, learners can monitor whether they understand the chemical process whereby carbon dioxide is converted into carbohydrates, for example, by self-testing leading to monitoring judgments. This activity also reflects monitoring in the learning phase, but with retrieval as the object of monitoring. However, monitoring is not limited to the learning (or preparation) phase but ideally continues in the testing phase when learners retrieve or transfer their knowledge, for example in an exam or application task. Here, learners can again monitor their processing. As described above, students can first strategically monitor their test taking behavior. For example, when taking a written exam on the topic of photosynthesis, they can monitor whether their chosen strategy to answer all questions in the predetermined order is leading to success or if they predict that they will run out of time without completing all the high-value open answer questions (monitoring strategy). Second, they can monitor their retrieval during task processing which may result in varying levels of confidence in their selected answers (monitoring judgments). As a result, they may then carefully check whether their chosen answer in a multiple choice item is the (only) correct one.

Overall, in this example, we can see different aspects of metacognitive monitoring: During the knowledge acquisition or learning phase as well as during the testing phase, monitoring can refer to processing and to retrieval. However, research so far has studied those aspects of metacognitive monitoring in a quite isolated manner and without considering their complex relationship. In addition, not only the object of monitoring (i.e., processing vs. retrieval) and the phases where monitoring occurs (learning phase vs. testing phase) can differ (Händel et al., 2023) but – according on the underlying theory or model – also the granularity can differ depending on the type of methodological procedure that is implemented to assess metacognitive monitoring (Rovers et al., 2019). That is, monitoring strategies are typically assessed on a quite general level across multiple tasks and monitoring judgments are assessed with reference to specific tasks or items within a task. It seems a worthwhile endeavor to assess various types of metacognitive monitoring within one sample and one learning context. To overcome the limitation of lacking information on the quality of monitoring strategies, recent procedures have been developed that ask students to reflect on the quality of their implemented monitoring strategies (Händel & Dresel, 2022). Further, diary studies (Eckerlein et al., 2019) and experience sampling methods (Nett et al., 2012) that target specific learning activities offer the opportunity to overcome the limitation that questionnaires often reflect rather trait-like learning behaviors.

1.2. Interplay of monitoring strategies and judgment accuracy

According to the model by Händel and Dresel (2022), monitoring strategies and monitoring judgments are reciprocally related: The use and quality of monitoring strategies should enhance judgment accuracy and vice versa. Back to the example with the exam on photosynthesis, this means that students who use strategies to monitor their learning, for example via self-testing or explaining photosynthesis content to others, might thereby get insights into their understanding and accordingly make more accurate judgments. Vice versa, students who are able to accurately judge their performance and detect knowledge gaps might not only invest effort to close these knowledge gaps but also use strategies to monitor whether they are successful regarding this purpose (see also Morphew, 2021).

First empirical evidence on the interplay of monitoring strategies and monitoring judgments stems from training studies, indicating that monitoring strategy training has (at least short-term) positive effects on judgment accuracy (de Bruin et al., 2017; Huff & Nietfeld, 2009; Nietfeld et al., 2006; Nietfeld & Schraw, 2002). Notably, the underlying processes are unclear, especially as monitoring strategy trainings were mostly combined with feedback, which might directly improve the accuracy of judgments (Follmer & Tise, 2021; Naujoks et al., 2022). Remarkably, studies that aim to disentangle the relationship between monitoring strategies and monitoring judgments found only weak correlations between the different components of metacognitive monitoring, i.e., monitoring strategies and judgment accuracy (Dörrenbächer-Ulrich et al., 2021; Dörrenbächer-Ulrich & Perels, 2023; Eberhart et al., in press; Golke et al., 2022; Händel & Dresel, 2022). However, the cross-sectional design of these studies limits insights regarding any causality.

Potential explanations for the low correlations are that monitoring strategies and monitoring judgments are usually assessed in different phases of the learning or testing process (Händel et al., 2023)—strategies during studying and judgments during testing—as well as on a different level of granularity (Rovers et al., 2019)—strategies referring to learning in general and judgments referring to specific tasks.

Monitoring judgments are typically assessed as *explicit or prompted judgments* in testing situations where students are explicitly asked how confident they are about their task solution (Barenberg & Dutke, 2018; Händel, de Bruin, et al., 2020; Jordano & Touron, 2018; Kruger & Dunning, 1999) and judgment accuracy is typically the indicator of interest in this branch of research (Schraw, 2009; Schraw et al., 2013). Collecting monitoring judgments during test has the advantage that the judgment is clearly related to a performance score and can be evaluated against a standard (Winne, 2017), e.g., via several accuracy scores (Schraw, 2009; Schraw et al., 2013). For example, sensitivity or specificity (Rutherford, 2017) as the two scores used in the current study, indicate the relative amount of accurately or inaccurately detected correct answers.

Still, those prompted judgments might not provide any information about naturally occurring monitoring processes or (spontaneous) judgments in the learning phase. In learning situations under realistic conditions – see the learner above preparing for an exam on the topic of photosynthesis – metacognitively skillful students should spontaneously generate judgments that influence their further learning behaviors or learning choices. However, assessment methods for spontaneous judgments like eye tracking methods (Roderer & Roebers, 2010), think-aloud protocols (Jordano & Touron, 2018; Schellings et al., 2013), uncertainty gestures (Dutemple et al., 2023), or behavioral coding (Bryce &

Whitebread, 2012) might influence natural monitoring behavior.

Overall, "there is little work addressing how frequently individuals spontaneously think about their task performance" (Jordano & Touron, 2018, p. 1272) and it is unclear whether the method of prompting judgments leads to monitoring processes that otherwise would not occur (Double & Birney, 2019). Hence, the question is whether *prompted judgments* resemble *spontaneous judgments* (Halamish & Undorf, 2020) and how they relate to performance (Halamish & Undorf, 2020; Jordano & Touron, 2018). Previous (lacking) correlations of monitoring strategies and judgment accuracy, thereby, might be due to lacking validity of prompted judgments. As such, an important contribution of the current study is the collection of monitoring strategies during studying, spontaneous monitoring judgments during studying and prompted monitoring judgments during test – all related to the same learning task.

1.3. Interindividual differences in monitoring strategies and monitoring judgments

Learners differ in the amount or quality of implemented monitoring strategies or in the accuracy of their metacognitive judgments. While individual differences in monitoring accuracy have been established for a long time (Maki, 1998), little consensus has been reached about definitive person-related predictors of these differences, with different studies collecting different person-related factors and producing mixed results (Golke et al., 2022; Händel, de Bruin, et al., 2020). Cognitive (e.g., prior knowledge), metacognitive (e.g., metacognitive knowledge), and demographic (e.g., gender) factors might contribute to both monitoring strategies and monitoring judgments.

To adequately monitor learning or performance in a specific topic, prior knowledge or domain knowledge seems necessary. Learners with high topic knowledge might have more cognitive resources available that they can invest in monitoring activities (Seufert et al., 2024). According to research findings, prior knowledge can support students in applying monitoring strategies (Hemmler & Ifenthaler, 2024; Taub et al., 2014) as well as in accurately judging their performance (Golke et al., 2022; Händel, de Bruin, et al., 2020; Kruger & Dunning, 1999; Mihalca & Mengelkamp, 2020; Prinz et al., 2020).

As a metacognitive variable, metacognitive knowledge (Flavell, 1979; Tarricone, 2011) has been discussed and studied in the past. Learners who have higher strategy knowledge about monitoring strategies, their use, and usefulness are more likely to implement those strategies (Karlen, 2016; Tuononen et al., 2023). Similarly, learners with higher knowledge about the individual strengths and weaknesses and task difficulties should be more accurate in their monitoring judgments (Dörrenbächer-Ulrich & Perels, 2023; Händel & Dresel, 2022; Terneusen et al., 2024). In a current meta-analysis on interventions for monitoring accuracy in problem solving, different intervention types moderated the training effect on monitoring accuracy, with interventions that targeted metacognitive knowledge showing the largest effect size (Janssen & Lazonder, 2024).

Furthermore, previous research found gender-specific patterns in monitoring strategies as well as judgment accuracy. The study by Golke et al. (2022) tested the contribution of various learner characteristics to situation-specific pre- and postdiction bias in a text comprehension task and observed, that gender was a consistent predictor. In this case, males and especially those with lower reading skills showed more bias (i.e., poorer monitoring accuracy). Händel, de Bruin, et al. (2020) found that female students showed lower sensitivity but higher specificity scores. That is, female students were better at detecting correct answers but worse in detecting incorrect answers. Overall, findings seem to be mixed and possibly related to self-concept (Händel, de Bruin, et al., 2020; Jansen et al., 2014) or self-efficacy (Golke et al., 2022), as indicated by studies with typical male or female domains. Namely, in domains considered typically male like mathematics of physics, male learners show more positive self-concepts than female students while in domains considered typically female like reading or psychology, female students report higher domain-specific self-concepts (Ertl et al., 2017; Franzen et al., 2022; Sax et al., 2015; Skaalvik & Skaalvik, 2004). Overall, in the context of interdependencies of monitoring strategies and monitoring judgments, it seems worthwhile to consider individual differences.

1.4. The current study

To advance our understanding of metacognitive monitoring, the current study takes an integrated approach by connecting research lines from monitoring strategies and metacognitive judgments and considers individual differences for both research lines. The study used a micro-longitudinal, crossed-lagged design to investigate how monitoring strategies and monitoring judgments change micro-genetically and influence each other reciprocally during an exam preparation phase. To overcome previous limitations, the study considered the following aspects:

Spontaneous vs. *prompted monitoring judgments*. Previous research on the interplay of monitoring strategies and monitoring judgments typically asked learners to explicitly provide monitoring judgments, on the assumption that they were capturing a naturally occurring monitoring process. However, it is possible that students would not have otherwise made a judgment spontaneously, for example, in a low stakes learning situation, when the learning situation is not highly important for them, when the learners are not motivated for the topic, or when the learners are distracted. Therefore, to investigate the relationship of monitoring strategies and monitoring judgments, the current study aimed to investigate spontaneous monitoring judgments, considering both, positive spontaneous judgments (i.e., learners positively evaluating their understanding) and negative spontaneous judgments (i.e., learners recognizing difficulties in learning or understanding) in addition to prompting explicit metacognitive judgments.

Phases of learning or testing. Previous research revealed lacking or only low correlations between monitoring strategies and judgment accuracy. A potential explanation is that the two types of monitoring are assessed in different phases of studying. Hence, in addition to the study of prompted judgments in the testing phase, this study aimed to study monitoring strategies and spontaneous metacognitive judgments within the same phase, namely the learning phase aiming to understand the relationship between the two.

Level of assessment – granularity. To investigate monitoring strategies and spontaneous metacognitive judgments at the same granularity level, the study aimed to each refer monitoring to the whole learning situation addressed in this study (i.e., daily preparation for an exam).

Reciprocal relations. Previous findings regarding the interplay of monitoring strategies and judgment accuracy were restricted to single learning situations within cross-sectional research designs, and therefore do not inform about how different components of metacognitive monitoring influence each other during the learning process, presumably in a reciprocal manner. Accordingly, little is known about micro-development and reciprocal relationships of monitoring strategies and judgments. Hence, the current study aimed to contribute to a better theoretical understanding of the interplay of the two via a micro-genetic diary study.

With these considerations, the current study investigated four hypotheses, listed below.

Individual differences: For both metacognitive monitoring strategies and judgments, individual differences are expected with regard to prior knowledge, metacognitive knowledge, and gender. However, given the lack of prior evidence, more differentiated hypotheses for the different types of monitoring measures were not formulated (namely, monitoring strategies, spontaneous judgment – positive and negative, as well as sensitivity and specificity as indicators for judgment accuracy).

H1a. GPA, prior subject performance, and metacognitive knowledge are positively related to metacognitive monitoring.

H1b. Metacognitive monitoring shows gender-specific patterns, related to students' self-concepts.

Relationship of spontaneous and prompted judgments: Positive relationships are assumed between spontaneous judgments and the accuracy of prompted judgments. That is, students who more strongly experience spontaneous judgments during studying should show higher judgment accuracy during testing.

H2. Experience of spontaneous judgments and accuracy of prompted judgments are related to each other.

Development of metacognitive monitoring: While it is reasonable to expect a change in how the various aspects of monitoring affect each other over time, given the novelty of this study design in this field, an undirected hypothesis regarding the development of the different types of metacognitive monitoring was studied.

H3. Monitoring strategies, spontaneous monitoring judgments, and accuracy of prompted judgments change during an intense exam preparation phase.

Reciprocal relationships: According to the model by Händel and Dresel (2022), we do expect to observe evidence of reciprocal relationships in this study. Namely, learners who implement monitoring strategies in high quality to monitor their learning behavior should gain more insight into their strengths and weaknesses and thus might report a stronger experience of (and more accurate) judgments accordingly. Conversely, more accurate judgments could lead to learners monitoring their learning behavior more closely. However, since this may depend on the type of judgment (not only spontaneous vs. prompted), but also on the level of knowledge (positive or negative judgment/accurate judgment of high or low knowledge), no strong hypotheses were formulated regarding how these relationships change across time.

H4. Monitoring strategies and monitoring judgments (i.e., spontaneous judgments as well as accuracy of prompted judgments) influence each other across time.

2. Method

2.1. Study design

The diary study took place in the week before a final course exam and was conducted using Unipark Questback ESF. The diary study consisted of one initial survey plus five measurements collected on consecutive days. It aimed to assess metacognitive monitoring from a process-based perspective in an intense learning and self-testing phase one week before an exam. Each daily measurement consisted of a short survey part as well as a knowledge test part. All variables collected in the study related to the daily knowledge test or the preparation for it. The procedure of each day was the same: Students first answered questionnaire items on their spontaneously generated metacognitive judgments during studying and on the quality of their monitoring strategies during studying. Second, students answered the knowledge test items, and provided metacognitive judgments for each knowledge test question (prompted item-specific judgments). In addition, one week before the daily testing session started, students participated in an initial survey and provided information on the semester they were enrolled in, their gender, indicators for their prior performance (GPA as well as their grade in the preceding lecture) as well as their metacognitive knowledge and their domain-specific self-concept.

All students enrolled in the course were invited to take part in the study via individual daily *E*-Mails. Each invitation link was valid for 24 h and each student could take part only once per day. Participation in the survey as well as the knowledge test was voluntary. Students were informed that the purpose of the survey was purely scientific, that no disadvantages could occur due to non-participation and that their data would be handled only using a pseudonym. Only students who provided consent for the study could take part in the knowledge tests and accompanying survey. The study was conducted fully in line with the Ethical Principles of Psychologists and Code of Conduct of the American Psychological Association (American Psychological Association, 2017) and in accordance with the Code of Ethics of the World Medical Association (2013) – declaration of Helsinki; an institutional ethics approval was not obtained. As an incentive for participation in the survey students received €2 for the pretest, €1 for each testing day and €3 extra if they participated in all five testing days.

2.2. Sample

Overall, $N = 273$ higher education students of an educational bachelor program participated in the initial survey and at least one of the five testing days. Participants' gender distribution was: 73.6 % female, 25.9 % male, 0.5 % non-binary with the high proportion of female students being typical for courses of this study program. On average, students were enrolled in the third semester ($M = 3.2$, $SD = 1.5$), that is, the second study year of their bachelors' studies. Students' final high school grade (GPA, ranging from 1 = *excellent* to 6 = *insufficient*) was $M = 2.3$, $SD = 0.6$. Table 1 provides an overview of the participation rates across the five measurements. As study participation was voluntary and possible only within a 24 h window each day, not all students participated in all testing days. In addition, missing values (omitted) were found in only very few cases, ranging from 0.1 % at Day 1 to 2.6 % missing values at Day 4.

2.3. Instruments

To assess monitoring in a manner that is as situation-specific as possible without disturbing learners in their natural learning processes as is it the case, for example, via think-aloud protocols (Schellings et al., 2013) or eye-tracking apparatus (Roderer & Roebers, 2010), we used daily and situation-specific self-reports. We took highest care to use economic and short questionnaire scales, usually with only two or three items. Other studies already provided evidence regarding the feasibility of such an approach (Loeffler et al., 2019). With long item batteries, students might not have been willing to participate in the study during the intense exam preparation phase.

Students were asked to answer questions regarding the quality of their implemented monitoring strategies and to report their experience of spontaneous monitoring judgments, each in relation to the preparation for the daily performance test. In addition, students provided item-specific metacognitive judgments regarding each item in the performance test.

2.3.1. Monitoring strategies

Quality of strategies to monitor learning. To overcome the limitation of lacking information on the quality (Wirth & Leutner, 2008) as well as on the situation-specific nature (Rovers et al., 2019) of monitoring strategies in popular instruments like the MLSQ (Pintrich et al., 1991), a situation-specific approach to assess the quality of strategies to monitor learning was implemented via the 3-items from Händel and Dresel (2022). Students first described in their own words how they had monitored their learning during the preparation of the daily performance test. Sample answers were: "I made notes of my mistakes"; "self-

Table 1
Participation Rates for 1 [min] to 5 [max] Testing Days.

Number of testing days students participated in	Proportion of students [%]	Cumulative values [%]
5	54.0	54.0
4	12.8	66.8
3	9.6	76.4
2	9.6	86.0
1	14.0	100.0

testing"; "I tried to explain the content in my own words".[1] Students then self-judged the quality of the implementation of their monitoring strategy: They answered three items regarding the correct and effective implementation of their personal strategy (target orientation, precision, and control). A sample item was "I made sure that I used the strategy correctly". The items were answered on a 6-point Likert scale (1 = *not true at all* to 6 = *absolutely true*). The final variable was computed as the mean score of the three items for each testing day (α = 0.91 to 0.96).

2.3.2. Monitoring judgments

Spontaneous monitoring judgments were assessed with four items – two items covered spontaneous judgments where the learners had a Eureka moment (so-called positive spontaneous judgments; sample item "At some point during studying, I spontaneously thought: 'now I've remembered it'") and two items covered spontaneous judgments where the learners had recognized that they did not yet understand or remember the learning content (so-called negative spontaneous judgment; sample item: "While studying, I suddenly realized that I still don't understand the content"). The items were answered on a 6-point Likert scale (1 = *not true at all* to 6 = *absolutely true*). The items were based on a think-aloud study (Schellings et al., 2013) as well as on a pilot study with originally eight items (four positive and four negative judgments). The resulting variables 'spontaneous positive judgment' (α = 0.77 to 0.93) and 'spontaneous negative judgment' (α = 0.63 to 0.88) were computed as the mean score of each the two items for each testing day.

Prompted item-specific monitoring judgments were assessed along with each answer option in the performance test (Händel & Dresel, 2022). Each of the five daily performance tests comprised eight multiple choice items each with four answer options. The items related to course content which had been taught throughout the semester and which was relevant to the final exam, but identical items did not appear on the final exam. Within each testing day, students had to decide true/false for 8*4 = 32 answer options. Students also provided a metacognitive judgment for each answer option. The item text was "Please indicate what you think about your answer. My answer was… (1 = *not correct at all* to 6 = *absolutely correct*)".

In line with previous studies, sensitivity and specificity scores were calculated as relative scores of judgment accuracy (Händel, Harder, et al., 2020; Schraw et al., 2013). To do so, answers 1 to 3 were coded as "incorrect" and answers 4 to 6 as "correct". Sensitivity indicates the relative amount of accurately detected correct answers while specificity indicates the relative amount of accurately detected incorrect answers. Sensitivity is calculated as the proportion of so-called hits (correct answers that were identified as such) divided by the sum of all correct answers. Vice versa, specificity is calculated as the proportion of correct rejections (incorrect answers detected as such) divided by the sum of all incorrect answers. The choice of sensitivity and specificity scores alongside a plethora of different relative and absolute scores for judgment accuracy is grounded in two deliberations. First, the two scores complement each other, and students can show high sensitivity as well as high specificity within the same test as sensitivity refers to the detection of correct items while specificity refers to the detection of incorrect items; see, for example, Rutherford (2017) who constated that sensitivity and specificity "… had different relations both between and within students provides further evidence that they represent separate metacognitive evaluation processes" (p. 236). Second, as students' spontaneous judgments were assessed separately for positive and negative judgments, those spontaneous judgments can each be set into relation with sensitivity (positive spontaneous judgments) and with specificity (negative spontaneous judgments).

2.3.3. Preconditions of monitoring

Prior performance. In addition to their GPA, students were asked to indicate their grade in the preceding psychology lecture (1 = *excellent* to 6 = *insufficient*).

Metacognitive knowledge. To assess metacognitive knowledge regarding monitoring strategies before the diary period started, a subset of the scenario-based knowledge test by Dörrenbächer-Ulrich et al. (2024) was implemented. Two scenarios were described in which a student is preparing for an exam. In each scenario, students are provided with six strategy options to deal with the situation. According to expert ratings, three of the given strategies are regarded as useful and three as less useful strategies. Study participants had to rate the six strategy options per scenario on a 4-point Likert scale of usefulness (1 = *not useful at all*; 4 = *very useful*). The scoring of the test was applied via pair comparisons of each two strategies, resulting in 18 pair comparisons and their respective credits.

Domain-specific self-concept. To further investigate potential gender differences, domain-specific self-concept was assessed via a 5-item scale by Schöne et al. (2012) adapted to the domain of psychology. A sample item was: "I am good at psychology". Internal consistency was good with Cronbach's α = 0.90.

2.4. Data analysis

The variables of interest for metacognitive monitoring in the current study were a) the quality of monitoring strategies in the learning phase, b) spontaneous positive judgments in the learning phase, c) spontaneous negative judgments in the learning phase, and, finally d), sensitivity and e) specificity as two relative scores reflecting accuracy of monitoring judgments in the testing phase.

To address H1a, Pearson's correlations with the variables from Day 1 were calculated. H1b was investigated via a MANOVA with gender as independent variable and all monitoring scores at Day 1 as dependent variables. The data from Day 1 was selected because it cannot be impacted by potential practice effects of providing repeated monitoring reports.[2] H2 was investigated via Pearson's correlations for each individual testing day. To study H3 and H4, random intercept cross-lagged panel models were calculated using the data set from Days 1–5 (RI-CLPM; Hamaker et al., 2015).[3] In four separate models, autoregressive paths as well as the cross-lagged paths of monitoring strategies and monitoring judgments were estimated (Fig. 1). In each of the four models, the quality of monitoring strategies was investigated on the one hand, and either (1) spontaneous positive judgments, (2) spontaneous negative judgments, (3) sensitivity, or (4) specificity on the other hand.

RI-CLPMs enable the separation of stable, person-related variances from within-person processes. This allows for a more accurate analysis of within-person relationships and processes over time (Hamaker et al., 2015). In these models, the random intercepts (RI_Strategies and RI_Judgments) capture the time-invariant stability while the autoregressive paths (α_J and α_S) account for the additional carry-over effects from one day to the other. To achieve simplicity in structural equation models, specific parameters were constrained: all factor loadings, both for trait and state variables, were fixed at 1. Additionally, the variances of state variables were set to remain stable over time. Auto-regressive and cross-lagged paths were fixed to be equal over time. Measurement errors of the manifest variables were permitted to correlate for corresponding variables.

3. Results

3.1. Descriptive statistics

Students' performance in the preceding lecture (prior performance)

[1] Students' answers were not further coded by the researchers.

[2] Additional analyses using the mean score of all testing days showed the same pattern of results.

[3] The estimator MLR was used to obtain robust estimates with missing data.

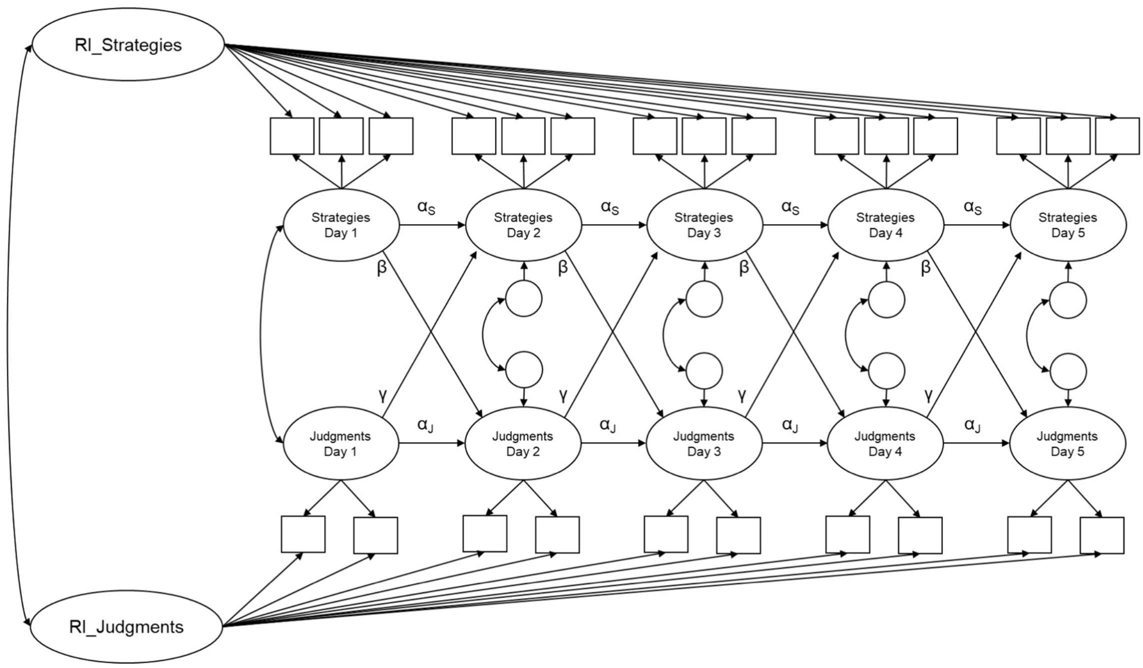

Fig. 1. Structure of the Estimated Random Intercept Cross-Lagged Panel Models (all Factor Loadings are Fixed to 1; Measurement Errors of the Manifest Variables are not Depicted, They Were Permitted to Correlate for Corresponding Variables).
Note. This model structure was tested for four different set of variables: each with monitoring strategies on the one hand and (1) spontaneous positive judgments, (2) spontaneous negative judgments, (3) sensitivity, and, finally (4) specificity on the other hand.
RI_Strategies = random intercept of the variable monitoring strategy quality; RI_Judgments = random intercept of the variables (1) spontaneous judgment positive, (2) spontaneous judgment negative, (3) sensitivity, and (4) specificity.
α_S and α_J display the autoregressive paths for the variable strategies or judgments (either spontaneous positive, negative or sensitivity or specificity).
β and γ display the cross-lagged paths.

was $M = 1.95$, $SD = 0.77$ (ranging from 1 to 6, where 1 reflects the best performance). Students' metacognitive knowledge (ranging from 0 to 18 points) was above the scale average ($M = 13.60$, $SD = 3.88$). Similarly, students' domain-specific self-concept (ranging from 1 to 6) was above average ($M = 3.98$, $SD = 0.73$). Descriptive statistics for the daily assessed variables of interest per testing day are shown in Table 2.

The students' practice test performance was good and reflects that the students were in the final stages of preparing for the exam. Students reported above average quality of their monitoring strategies as well as above average experience of spontaneous judgments – in similar amount for spontaneous positive and negative judgments. For sensitivity and specificity (calculated from their prompted item-specific judgments), results were different. Students' sensitivity, that is, the ability to accurately detect correct answers, was high, while their specificity, that is, the ability to accurately detect incorrect answers was rather low. Overall, the level of each dependent variable remained quite stable across the five consecutive testing days.

3.2. Interindividual differences

Surprisingly, none of the monitoring scores from Day 1 significantly correlated with GPA or metacognitive knowledge (H1a, see Table 3). In contrast, higher prior performance scores were associated with higher sensitivity and lower specificity of item-specific monitoring judgments. That is, students with higher performance in the lecture exam of the previous semester were better able to detect correctly solved items but less able to detect incorrectly solved items. These correlations remain significant even when performance at Day 1 was statistically controlled for and, thereby, seems not only to be driven by the performance scores.

When investigating gender differences for the monitoring scores at the first testing day (H1b), again, significant effects were only found for sensitivity ($F(1, 169) = 4.72$, $p = .031$, eta$^2 = 0.03$) and specificity ($F(1, 169) = 7.61$, $p = .006$, eta$^2 = 0.04$) with female students ($M = 0.78$, $SD = 0.21$) being better able to detect correctly solved items (i.e., higher sensitivity) than male students ($M = 0.70$, $SD = 0.24$) but being less able to detect incorrectly solved items (i.e., lower specificity; $M = 0.22$, $SD =$

Table 2
Means and Standard Deviations (in Brackets) for all Monitoring Scores and Each Testing Day.

Daily variables	Range	Day 1	Day 2	Day 3	Day 4	Day 5
Practice test performance	[0–100 %]	68.4 (8.9)	76.2 (10.5)	75.4 (10.8)	75.7 (11.6)	69.4 (11.2)
Quality of monitoring strategies	[1–6]	4.08 (1.06)	3.99 (1.26)	4.15 (1.16)	4.14 (1.16)	4.17 (1.27)
Spontaneous positive judgment	[1–6]	3.68 (0.96)	3.68 (1.11)	3.74 (1.08)	3.85 (1.12)	3.88 (1.18)
Spontaneous negative judgment	[1–6]	3.69 (0.76)	3.69 (0.80)	3.58 (0.76)	3.63 (0.78)	3.53 (0.78)
Prompted monitoring judgment	[1–6]	4.25 (0.72)	4.35 (0.71)	4.28 (0.77)	4.33 (0.72)	4.21 (0.73)
Sensitivity	[0–1]	0.76 (0.22)	0.80 (0.20)	0.76 (0.24)	0.78 (0.21)	0.78 (0.23)
Specificity	[0–1]	0.26 (0.22)	0.33 (0.30)	0.25 (0.25)	0.26 (0.26)	0.26 (0.25)
N		201	194	188	195	180

Note. Spontaneous positive and negative judgments represent the experience of spontaneous monitoring.
The prompted monitoring judgment indicates the judgment regarding specific items in the daily performance tests.
Sensitivity and specificity are accuracy measures of the prompted judgments.

Table 3
Bivariate Correlations between GPA, Prior Performance, Metacognitive Knowledge, and Monitoring at Day 1.

	Monitoring at Day 1				
Potential antecedents	Monitoring strategy quality	Spontaneous judgment positive	Spontaneous judgment negative	Sensitivity	Specificity
GPA	0.09	−0.03	0.06	0.08	−0.04
Prior performance	0.06	0.01	−0.13	0.16*	−0.20*
Metacognitive knowledge	−0.06	−0.10	0.01	0.05	−0.05

Note. GPA and prior performance are recoded so that higher scores indicate better performance.
* $p < .05$.

0.21) than their male counterparts ($M = 0.33$, $SD = 0.24$). These gender differences remained significant and with the same effect sizes even when domain-specific self-concept was controlled for.

3.3. Correlations between spontaneous and prompted judgments

Table 4 shows the correlations between spontaneous positive judgments and sensitivity for each testing day as well as between spontaneous negative judgments and specificity for each testing day (addressing H2). Overall, correlations were either non-significant or small, indicating only small correlations between the experience of spontaneous judgments and the accuracy of prompted judgments. If significant, students who reported a stronger experience of spontaneous positive judgments showed higher sensitivity (being better able to detect correct items) and students who reported a stronger experience of negative spontaneous judgments showed higher specificity (being better able to detect incorrect items).

3.4. Interrelationship between monitoring strategy quality and monitoring judgments

Finally, RI-CLPMs were applied to investigate H3 and H4. Results in Table 5 indicate that the models for all investigated variables (each quality of monitoring strategies as one variable and one type of judgment or judgment accuracy as the other variable) showed a good fit.

For the quality of monitoring strategies, the variances of the random intercepts, representing the time-invariant stability, were non-significant (cf., Supplement 2). However, the autoregressive paths indicate significant carry-over effects from one day to the other. In contrast, both types of spontaneous judgments as well as sensitivity and specificity provide high time-invariant stability (see variances of RI_Judgments in Table 1, Supplement 2) with no additional carry over effects from one day to the next. Regarding the cross-lagged paths (H4), a higher quality of monitoring strategies led to stronger experience of spontaneous judgments, both for positive spontaneous judgments as well as for negative spontaneous judgments (see Fig. 2, upper part). The reverse path from spontaneous judgments to quality of monitoring strategies was not statistically significant ($p = .056$). Finally, no statistically significant cross-lagged paths were observed between quality of monitoring strategies and the accuracy scores of prompted monitoring judgments (see Fig. 2, lower part). Still, there was a significant relation between the stable trait-like components of quality of monitoring strategies and sensitivity (but not for specificity; see RI_Strategies × RI_Judgments).

Table 5
Model Fit Indices of the four RI-CLPMs.

Strategy	Quality of monitoring strategies			
Type of judgment	Spontaneous pos.	Spontaneous neg.	Sensitivity	Specificity
CFI	0.963	0.940	0.958	0.966
TLI	0.959	0.933	0.951	0.961
RMSEA	0.044	0.054	0.055	0.046
χ^2	398.37, $df =$ 269	463.14, $df =$ 269	286.67, $df =$ 164	249.00, $df =$ 164

4. Discussion

The current study aimed to contribute to our understanding of the complex interrelationship of different facets of metacognitive monitoring by considering approaches from two strands of research – monitoring strategies from the perspective of self-regulated learning and monitoring judgments from the perspective of metacognition. In addition, by considering both spontaneous and prompted judgments, the study adopts a comprehensive approach, integrating perspectives from cognitive and educational psychology. Monitoring strategy quality and monitoring judgments were examined simultaneously in a 5-day diary study during the ecologically valid learning situation of exam preparation including several self-testing opportunities. First, besides the accuracy of explicit judgments prompted during test completion, the current study asked students to report on the spontaneous generation of monitoring judgments during studying. In addition, to overcome limitations of previous studies that had provided evidence on the interplay of monitoring strategies and monitoring judgments for single testing situations, this study used a micro-genetic longitudinal design. Finally, the use of RI-CLPM allowed us to study the reciprocal relationship as well as intraindividual personal stability and changes over time in monitoring strategy quality and monitoring judgments.

Data were analyzed to address four hypotheses. The results regarding the first hypothesis on interindividual differences in monitoring strategy quality and monitoring judgments indicated that prior performance and

Table 4
Correlations Between Spontaneous Judgments and Accuracy of Prompted Judgments.

Positive / negative spontaneous judgment	Sensitivity / Specificity				
	Day 1	Day 2	Day 3	Day 4	Day 5
Day 1	−0.01 / 0.24**				
Day 2		0.18* / 0.06			
Day 3			0.20** / −0.08		
Day 4				0.11 /−0.04	
Day 5					0.08 / −0.12

Note. The first value per table cell displays the correlation coefficient between the positive spontaneous judgment and sensitivity. The second value per table cell displays the correlation coefficient between the negative spontaneous judgment and specificity.
** $p < .01$.
* $p < .05$.

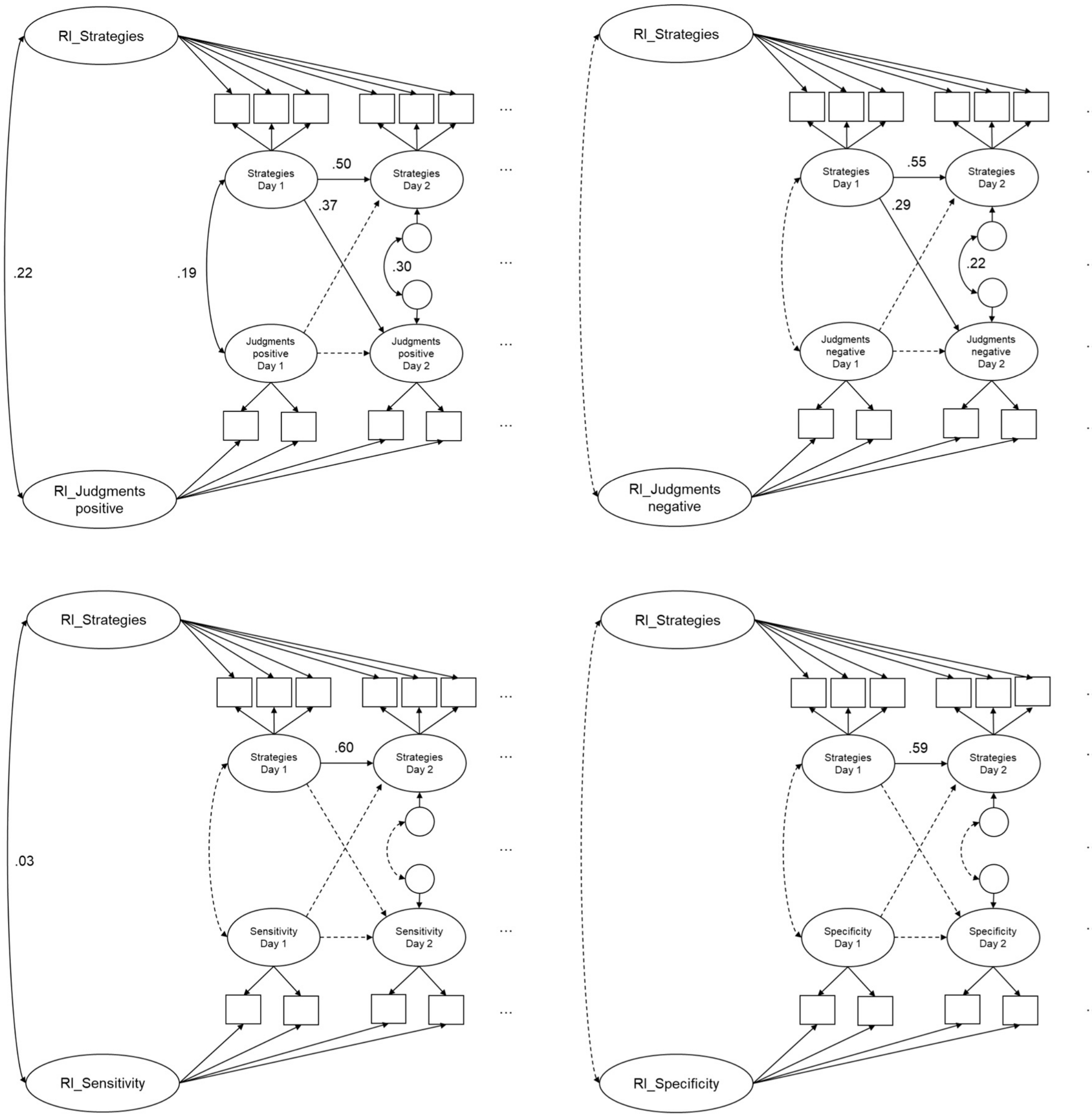

Fig. 2. Overview of Results for all Four Models.
Note. Where path coefficients are given, they are significant ($p < .05$); non-significant paths are displayed via dotted lines.
RI_Strategies = random intercept of the variable monitoring strategy quality; RI_Judgments positive = random intercept of the variable spontaneous judgment positive; RI_Judgments negative = random intercept of the variable spontaneous judgment negative; RI_Sensitivity = random intercept of the variable sensitivity (accuracy measure of the prompted judgments); RI_Specificity = random intercept of the variable specificity (accuracy measure of the prompted judgments).
The three dots in the right of each panel reflect Days 3–5.

gender were only associated with the accuracy of prompted monitoring judgments but not with spontaneous judgments during studying. The second hypothesis concerned the association between spontaneous monitoring judgments and the accuracy of prompted monitoring judgments and the evidence indicated a few weak positive associations across the five testing days. Thirdly, the quality of monitoring strategies showed little time-invariant stability but high carry-over effects indicating that the quality of monitoring strategies transferred from day to day. In contrast, the formation of spontaneous judgments as well as the accuracy of prompted judgments were rather stable across the five testing days without any additional carry-over effects. Lastly, the cross-lagged analyses that investigated the reciprocal relationship between the quality of monitoring strategies and spontaneous as well as the accuracy of prompted monitoring judgments indicated that the quality of monitoring strategies predicted subsequent formations of spontaneous monitoring judgments.

The current study, thereby, confirms previous propositions regarding the distinction between different phases and objects of monitoring (Händel et al., 2023). Furthermore, it extends the field of research by pointing to the relevance of distinguishing between the generation of spontaneous judgments and prompted judgments (Jordano & Touron, 2018).

4.1. *Prior performance and gender are associated with individual differences in monitoring judgment accuracy*

Nearly no correlations of GPA, prior performance, or metacognitive knowledge with either monitoring strategy quality or monitoring judgments (spontaneous or prompted) were observed. Only for the accuracy of prompted judgments, prior performance and gender were relevant factors. This is inconsistent with previous findings from Händel and Dresel (2022) or Golke et al. (2022). For GPA and prior performance this could be because students' high school grade and their grade in the preceding lecture are as such is not very specifically related to the knowledge tests which were the focus of this study. Partly consistent with Golke et al. (2022), interindividual differences regarding gender were found for the two monitoring judgment accuracy scores, even after controlling for domain-specific self-concept. In detail, female students showed higher sensitivity but lower specificity than their male counterparts. The direction of the gender effect is contrary to previous studies with other domains. For example, in a study using mathematics tasks, the pattern of results was reverse: Female students had lower sensitivity but higher specificity scores (Händel, de Bruin, et al., 2020). When investigating absolute monitoring accuracy in earlier studies (e.g., via bias or absolute accuracy), female students were often shown to be more cautious in estimating their performance (de Bruin et al., 2017). However, female students being less confident does not necessarily lead to lower judgment accuracy, as indicated by a study with intelligence tests (Ariel et al., 2018). Interestingly, domain-specific self-concept (Franzen et al., 2022) could not explain the gender effects, which is in line with bias regarding factual knowledge questions but not with bias for inference questions in the study by Golke et al. (2022). Hence, it remains a question for future research, why relative accuracy scores show specific gender-patterns in different domains and how motivational variables influence these patterns.

4.2. Spontaneous monitoring judgments and the accuracy of prompted monitoring judgments are weakly and inconsistently associated

Students in an intense exam preparation week reported to spontaneously generate monitoring judgments – both positive judgments (e.g., having a Eureka moment) and negative judgments (recognizing difficulties of understanding). However, the experience of those spontaneous judgments was not correlated or showed only small correlations with the accuracy of prompted judgments (measured via sensitivity or specificity). This result seems even more surprising than the low or lacking correlations of monitoring strategies and judgment accuracy in previous studies (Dörrenbächer-Ulrich et al., 2021;Händel et al., 2023; Händel & Dresel, 2022). Namely, while the relationship between metacognitive judgments and metacognitive strategy quality may be weaker simply because these involve different aspects of metacognitive monitoring (judgments vs. strategies), the investigation of the relationship between spontaneous and prompted judgments concerns metacognitive judgments in both cases. Hence, the low correlations shows the relevance of distinguishing spontaneous from prompted judgments (Halamish & Undorf, 2020). Still, it needs to be mentioned that the approach to assessing spontaneous judgments via self-reports was novel and not comparable to earlier studies that implemented laborious approaches like think-aloud protocols or behavioral measures (Bryce & Whitebread, 2012; Gidalevich & Kramarski, 2017; Greene et al., 2018). Assuming the findings will be replicated in future studies, it may require a reinterpretation of previous research on prompted metacognitive judgments as those judgments prompted in laboratory or testing situations might not display monitoring as it occurs naturally during studying. Consequently, the question arises to what extent the positive effects of training prompted judgments plus feedback as investigated in a couple of studies (Follmer & Tise, 2021; Miller & Geraci, 2011; Naujoks et al., 2022; Nietfeld et al., 2006) also transfer to naturally occurring monitoring during the learning process.

4.3. Spontaneous and prompted monitoring judgments are consistent within individuals during exam preparation, quality of monitoring strategies is not

Autoregressive paths in the RI-CLPM were significant for the quality of monitoring strategies, but not for monitoring judgments – neither the spontaneous ones, nor the accuracy scores of prompted judgments. This means that the time-invariant stability was rather low for the quality of monitoring strategies and that the quality of monitoring strategies transferred from day to day. A potential explanation might be that students probably adapted their effort in monitoring strategies in dependence of remaining days to the exam (Capelle et al., 2023). In contrast, spontaneous judgments as well as sensitivity and specificity were quite stable over the 5-day assessment period. That is, students seemed to have an individual habit in building spontaneous judgments or an individual competence in accurately detecting correct or incorrect items, which would be in line with the general monitoring skill hypothesis (Schraw & Nietfeld, 1998; Stankov et al., 2012). Since the spontaneous and the accuracy of prompted judgments were hardly correlated (see above), this result is remarkable because it shows that the same patterns can be found in both types of judgments across an intense exam preparation phase. Despite the differences in the time of assessment (learning vs. testing phase) and type of judgment (spontaneous vs. prompted), there seem to be commonalities here, which in turn distinguish the monitoring judgments from the quality of monitoring strategies.

4.4. The use of high-quality monitoring strategies leads to a stronger experience of spontaneous monitoring judgments

The study provides further evidence on the relationship between the quality of monitoring strategies and monitoring judgments. Like previous research that found only small correlations between monitoring strategies and judgment accuracy (Händel et al., 2023; Händel & Dresel, 2022), the stable trait-like component of monitoring strategy quality was correlated with only one of two scores for judgment accuracy, namely sensitivity. The current study confirmed and broadened previous research by investigating cross-lagged effects between the quality of monitoring strategies and two scores for judgment accuracy (namely, sensitivity and specificity) but found no significant cross-lagged paths. According to the study results, learners who use monitoring strategies in high quality during studying are not necessarily more accurate in their monitoring judgments regarding a performance test or vice versa. As data of the current study related to a quite intense and condensed preparation phase, they are consistent with the interpretation that previously found effects of strategy training on judgment accuracy might actually be driven by the feedback provided (Follmer & Tise, 2021).

In contrast, when it comes to the relationship of monitoring strategies quality and spontaneous monitoring judgments, the study revealed dependencies; namely, students who implemented monitoring strategies in high quality reported a stronger experience of spontaneous judgments – not only positive ones but also negative ones. The reverse path, however, was not significant. That is, spontaneous judgments did not lead to higher quality monitoring strategies. These results help to understand previous correlative results regarding the relationship between monitoring strategy quality and spontaneous judgments (Schellings et al., 2013) and contradict the idea of a reciprocal relationship (Händel & Dresel, 2022). For example, learners who implemented the monitoring strategy of self-explaining the lecture topic in their own words in high quality might have consequently generated a spontaneous monitoring judgment – either because they recognized that it was difficult (negative spontaneous judgment) or that it worked out well (positive spontaneous judgment). As can be easily imagined, students who implemented the monitoring strategy in low quality (e.g., too quickly looking up the content on the lecture transcript when self-testing) might have generated fewer spontaneous judgments than students who implemented the monitoring strategy in high quality (e.g., critically noticing knowledge gaps during self-testing). Transferred to training situations in educational settings, the study results suggest that supporting students in metacognitive processes during learning via the training of monitoring strategies might lead to stronger experience of spontaneous judgments.

Potential reasons for the lacking association between monitoring strategy quality and monitoring judgment accuracy might be methodological in nature. This may be due to the fact that monitoring strategies and judgments are implemented in different phases (Händel et al., 2023), or due to different task references—judgments were item-specific (Winne, 2017) while the use of monitoring strategies might be more general. As a result, despite our efforts to keep the reference point of the assessment as identical as possible, the two may still reflect different level of granularity (Rovers et al., 2019). In addition, the quality of monitoring strategies was assessed through self-judgment and therefore represents a subjective measure, whereas sensitivity and specificity are objective indicators of judgment accuracy. Further, it is possible that students did not naturally think about their item solution and only did so when they were explicitly asked to provide a monitoring judgment (Halamish & Undorf, 2020) whereas the application of monitoring strategies was self-initiated.

4.5. Limitations and implications

The current study broadens the field by assessing both spontaneous and prompted metacognitive judgments and by adopting a longitudinal study design. However, due to the implementation of the study into an ecologically valid situation of daily practice tests before a final lecture exam, the study could not control for some potential limiting factors. First, the assessment of spontaneous judgments was novel, restricted to self-reports and only piloted in one pre-study. Hence, it is not comparable to the think-aloud studies (Gidalevich & Kramarski, 2017; Greene et al., 2018; Jordano & Touron, 2018; Schellings et al., 2013) that assess spontaneous judgments during the learning process. Still, it had the advantage that it is economic in use and did not disturb the learning process. Second, in contrast to the prompted judgments that were item-specific and could be evaluated to given standards (test performance), the spontaneous judgments were only evaluated regarding the strength with which they were experienced but not to their adequacy or accuracy. Third, despite the tests implemented in the current study being of high relevance (i.e., same topic as in the upcoming exam; only a few days before the upcoming exam), and students seemed to take great care to use monitoring strategies, the tests were non-graded. As such, these findings should not be generalized to high stakes testing before being replicated in the context of graded exams. Finally, the aim of the study was to model changes in metacognitive monitoring over time within a very concrete 5-day-learning phase. Still, the chosen design with daily practice tests might serve as an intervention itself. For example, earlier studies showed that students who practice judgments may develop in their judgment accuracy (Foster et al., 2017; Hacker et al., 2000; Händel, Harder, et al., 2020). Hence, it might be possible that the self-report of spontaneous judgments generates more awareness and might influence monitoring judgments during the practice tests.

Overall, the study results point to the need to theoretically and empirically distinguish not only monitoring strategies and monitoring judgments but also the type of judgments (spontaneous vs. prompted). The latter is especially important for studies aiming at supporting students in their judgment accuracy because prompted judgments might not reflect students' natural monitoring behavior. As an implication for research, we suggest to not only use different (prompted) judgments (Händel, de Bruin, et al., 2020; Schraw, 1994) and different accuracy scores (Schraw et al., 2013) but to consider spontaneous judgments and the cues (Kelemen, 2000; Oudman et al., 2018) that they are based on. Interestingly, monitoring strategy quality and spontaneous monitoring judgments at the start of a 5-day learning phase before the exam showed nearly no individual differences regarding GPA, prior performance, metacognitive knowledge, or gender, which contrasts with earlier studies that used practice tests over the course of a semester. Only for sensitivity and specificity, prior knowledge in the domain under investigation as well as gender differences were found to be related to metacognitive monitoring. It seems that such a dense learning phase under pressure leads to typical learning habits (Capelle et al., 2023). Hence, future studies need to replicate those findings and consider motivational variables (Bartels & Magun-Jackson, 2009).

5. Conclusion

The current study provides first longitudinal evidence on the complex relationship of students' metacognitive monitoring processes during studying and testing. It reveals that prompted judgments, typically implemented in educational psychology research, are distinct from spontaneous judgments generated by students during their learning, and that only the latter were influenced by the quality of students' monitoring strategies. This is an important finding for the development of training materials aiming at supporting students in their monitoring processes.

CRediT authorship contribution statement

Marion Händel: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ulrike Nett:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Donna Bryce:** Writing – review & editing. **Markus Dresel:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2025.102748.

References

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code.

Ariel, R., Lembeck, N. A., Moffat, S., & Hertzog, C. (2018). Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence, 70*, 42–51. https://doi.org/10.1016/j.intell.2018.08.001

Baars, M. (2023). Understanding and supporting metacognitive monitoring judgments – A commentary. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 55*(2–3), 147–154. https://doi.org/10.1026/0049-8637/a000282

Barenberg, J., & Dutke, S. (2018). Testing and metacognition: Retrieval practise effects on metacognitive monitoring in learning from text. *Memory, 27*, 269–279. https://doi.org/10.1080/09658211.2018.1506481

Bartels, J. M., & Magun-Jackson, S. (2009). Approach–avoidance motivation and metacognitive self-regulation: The role of need for achievement and fear of failure. *Learning and Individual Differences, 19*, 459–463. https://doi.org/10.1016/j.lindif.2009.03.008

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

Capelle, J. D., Senker, K., Fries, S., & Grund, A. (2023). Deadlines make you productive, but what do they do to your motivation? Trajectories in quantity and quality of motivation and study activities among university students as exams approach. *Frontiers in Psychology, 14*, 1224533. https://doi.org/10.3389/fpsyg.2023.1224533

de Bruin, A. B. H., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning, 12*, 21–43. https://doi.org/10.1007/s11409-016-9159-5

Dörrenbächer-Ulrich, L., & Perels, F. (2023). Metacognitive judgment skills and the metacognitive component of self-regulated learning: A person-oriented, multimethod approach. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 36–48*. https://doi.org/10.1026/0049-8637/a000274

Dörrenbächer-Ulrich, L., Sparfeldt, J. R., & Perels, F. (2024). Knowing how to learn: Development and validation of the strategy knowledge test for self-regulated learning (SKT-SRL) for college students. *Metacognition and Learning, 19*(2), 1–45. https://doi.org/10.1007/s11409-024-09379-w

Dörrenbächer-Ulrich, L., Weißenfels, M., Russer, L., & Perels, F. (2021). Multimethod assessment of self-regulated learning in college students: Different methods for different components? *Instructional Science, 49*, 137–163. https://doi.org/10.1007/s11251-020-09533-2

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology, 10*, 2755. https://doi.org/10.3389/fpsyg.2019.02755

Dutemple, E., Hakimi, H., & Poulin-Dubois, D. (2023). Do I know what they know? Linking metacognition, theory of mind, and selective social learning. *Journal of*

Experimental Child Psychology, 227, Article 105572. https://doi.org/10.1016/j.jecp.2022.105572

Eberhart, J., Murayama, K., Sakaki, M., & Bryce, D. (in press). Metacognitive monitoring in early elementary school-aged children: Task dependency in monitoring judgements, task consistency in monitoring behaviours. Cognitive Development.

Eckerlein, N., Roth, A., Engelschalk, T., Steuer, G., Schmitz, B., & Dresel, M. (2019). The role of motivational regulation in exam preparation: Results from a standardized diary study. *Frontiers in Psychology, 10*, 81. https://doi.org/10.3389/fpsyg.2019.00081

Ertl, B., Luttenberger, S., & Paechter, M. (2017). The impact of gender stereotypes on the self-concept of female students in STEM subjects with an under-representation of females. *Frontiers in Psychology, 8*, 703. https://doi.org/10.3389/fpsyg.2017.00703

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

Follmer, D. J., & Tise, J. (2021). Across-task relations among monitoring judgments: Differential effects of item feedback on monitoring bias during reading. *Learning and Individual Differences, 88*, Article 102007. https://doi.org/10.1016/j.lindif.2021.102007

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

Franzen, P., Arens, A. K., Greiff, S., & Niepel, C. (2022). Student profiles of self-concept and interest in four domains: A latent transition analysis. *Learning and Individual Differences, 95*, Article 102139. https://doi.org/10.1016/j.lindif.2022.102139

Gidalevich, S., & Kramarski, B. (2017). Guidance for metacognitive judgments: A thinking-aloud analysis in math problem solving. *Hellenic Journal of Psychology, 14*(2), 83–113.

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, 2405–2450*. https://doi.org/10.1007/s10648-022-09687-0

Greene, J. A., Deekens, V. M., Copeland, D. Z., & Yu, S. (2018). Capturing and modeling self-regulated learning using think-aloud protocols. In D. H. Schunk, & J. A. Greene (Eds.), *Educational psychology handbook series. Handbook of self-regulation of learning and performance* (pp. 323–337). Routledge.

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

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), 304–305. https://doi.org/10.1027/2151-2604/a000429

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*, 102–116. https://doi.org/10.1037/a0038889

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

Händel, M., & Dresel, M. (2022). Structure, relationship, and determinants of monitoring strategies and judgment accuracy. An integrated model and evidence from two studies. *Learning and Individual Differences, 100*, Article 102229. https://doi.org/10.1016/j.lindif.2022.102229

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction, 65*. https://doi.org/10.1016/j.learninstruc.2019.101245

Händel, M., Naujoks-Schober, N., & Dresel, M. (2023). Metacognitive monitoring via strategies and judgments: Different phases, different objects. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 67–76*. https://doi.org/10.1026/0049-8637/a000273

Hemmler, Y. M., & Ifenthaler, D. (2024). Self-regulated learning strategies in continuing education: A systematic review and meta-analysis. *Educational Research Review, 45*, Article 100629. https://doi.org/10.1016/j.edurev.2024.100629

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*, 161–176. https://doi.org/10.1007/s11409-009-9042-8

Jansen, M., Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences, 30*, 11–21. https://doi.org/10.1016/j.lindif.2013.12.003

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

Jordano, M. L., & Touron, D. R. (2018). How often are thoughts metacognitive? Findings from research on self-regulated learning, think-aloud protocols, and mind-wandering. *Psychonomic Bulletin & Review, 25*(4), 1269–1286. https://doi.org/10.3758/s13423-018-1490-1

Karlen, Y. (2016). Differences in students' metacognitive strategy knowledge, motivation, and strategy use: A typology of self-regulated learners. *The Journal of Educational Research, 109*, 253–265. https://doi.org/10.1080/00220671.2014.942895

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92*, 800–810. https://doi.org/10.1037//0022-0663.92.4.800

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lan, W. (2005). Self-monitoring and its relationship with educational level and task importance. *Educational Psychology, 25*(1), 109–127. https://doi.org/10.1080/0144341042000294921

Loeffler, S. N., Bohner, A., Stumpp, J., Limberger, M. F., & Gidion, G. (2019). Investigating and fostering self-regulated learning in higher education using interactive ambulatory assessment. *Learning and Individual Differences, 71*, 43–57. https://doi.org/10.1016/j.lindif.2019.03.006

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). Routledge. https://doi.org/10.4324/9781410602350-13.

Mihalca, L., & Mengelkamp, C. (2020). Effects of induced levels of prior knowledge on monitoring accuracy and performance when learning from self-regulated problem solving. *Journal of Educational Psychology, 112*(4), 795–810. https://doi.org/10.1037/edu0000389

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314. https://doi.org/10.1007/s11409-011-9083-7

Morphew, J. W. (2021). Changes in metacognitive monitoring accuracy in an introductory physics course. *Metacognition and Learning, 16*(1), 89–111. https://doi.org/10.1007/s11409-020-09239-3

Naujoks, N., Harder, B., & Händel, M. (2022). Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. *Metacognition and Learning, 17*, 479–498. https://doi.org/10.1007/s11409-022-09295-x

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Vol. 26. Psychology of learning and motivation* (pp. 125–173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5.

Nett, U. E., Goetz, T., Hall, N. C., & Frenzel, A. C. (2012). Metacognitive strategies and test performance: An experience sampling analysis of students' learning behavior. *Education Research International, 1–16*. https://doi.org/10.1155/2012/958319

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159–179. https://doi.org/10.1007/s10409-006-9595-6

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131–142.

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*.

Prinz, A., Golke, S., & Wittwer, J. (2020). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31*, Article 100358. https://doi.org/10.1016/j.edurev.2020.100358

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

Rovers, S. F. E., Clarebout, G., Savelberg, H. H. C. M., de Bruin, A. B. H., & van Merriënboer, J. J. G. (2019). Granularity matters: Comparing different ways of measuring self-regulated learning. *Metacognition and Learning, 14*(1), 1–19. https://doi.org/10.1007/s11409-019-09188-6

Rutherford, T. (2017). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226–237. https://doi.org/10.1016/j.cedpsych.2017.03.001

Sax, L. J., Kanny, M. A., Riggers-Piehl, T. A., Whang, H., & Paulson, L. N. (2015). "But I'm not good at math": The changing salience of mathematical self-concept in shaping women's and men's STEM aspirations. *Research in Higher Education, 56*(8), 813–842. https://doi.org/10.1007/s11162-015-9375-x

Schellings, G. L. M., Van Hout-Wolters, B. H. A. M., Veenman, M. V. J., & Meijer, J. (2013). Assessing metacognitive activities: The in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education, 28*(3), 963–990. https://doi.org/10.1007/s10212-012-0149-y

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2012). *Skalen zur Erfassung des schulischen Selbstkonzepts (SESSKO) [Scales for assessing the school self-concept (SESSKO)]*. Hogrefe.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19*, 143–154. https://doi.org/10.1006/ceps.1994.1013

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415–429). Routledge.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007

Schraw, G., & Nietfeld, J. L. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90*, 236–248. https://doi.org/10.1037/0022-0663.90.2.236

Seufert, T., Hamm, V., Vogt, A., & Riemer, V. (2024). The interplay of cognitive load, learners' resources and self-regulation. *Educational Psychology Review, 36*(50). https://doi.org/10.1007/s10648-024-09890-1

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50*(3/4), 241–252. https://doi.org/10.1023/B:SERS.0000015555.40976.e6

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

Tarricone, P. (2011). *The taxonomy of metacognition*. Psychology Press.

Taub, M., Azevedo, R., Bouchet, F., & Khosravifar, B. (2014). Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Computers in Human Behavior, 39*, 356–367. https://doi.org/10.1016/j.chb.2014.07.018

Terneusen, A., Quaedflieg, C., Van Heugten, C., Ponds, R., & Winkens, I. (2024). The many facets of metacognition: Comparing multiple measures of metacognition in healthy individuals. *Metacognition and Learning, 19*(1), 53–63. https://doi.org/10.1007/s11409-023-09350-1

Tuononen, T., Hyytinen, H., Räisänen, M., Hailikari, T., & Parpala, A. (2023). Metacognitive awareness in relation to university students' learning profiles. *Metacognition and Learning, 18*(1), 37–54. https://doi.org/10.1007/s11409-022-09314-x

Winne, P. H. (2017). Cognition and metacognition within self-regulated learning. In P. A. Alexander, D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance*. Routledge.

Winne, P. H., & Hadwin, A. (2008). The weave of motivation and self-regulated learning. In D. Schunk, & B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Taylor & Francis.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie, 216*, 102–110. https://doi.org/10.1027/0044-3409.216.2.102

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Zimmerman, B. J. (1998). Academic studying and the development of personal skill: A self-regulatory perspective. *Educational Psychologist, 33*, 76–86. https://doi.org/10.1207/s15326985ep3302&3_3

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299–315). Routledge.