Educational Psychology Review (2024) 36:96
https://doi.org/10.1007/s10648-024-09936-4

META-ANALYSIS

Meta-analysis of Interventions for Monitoring Accuracy in Problem Solving

Noortje Janssen[1] · Ard W. Lazonder[1]

Accepted: 7 August 2024

Abstract

Accurate monitoring of performance in problem-solving tasks is an important prerequisite for students' future academic success. A wide variety of interventions aiming to enhance students' monitoring accuracy have been developed, but their effectiveness is not apparent from the individual studies in which they have been examined. This meta-analysis classified these interventions in terms of how they target students' monitoring and investigated their relative effectiveness to improve monitoring accuracy in problem-solving tasks. Findings across the 35 included studies indicated that all interventions combined have a small positive effect on students' monitoring accuracy ($g = 0.25$). Intervention type moderated the findings. Interventions on the whole task, metacognitive knowledge, and external standards improved monitoring accuracy. On the other hand, interventions targeting the timing of metacognitive judgment negatively impacted monitoring accuracy and significantly differed from all other interventions. Exploratory moderator analyses of study features indicated that secondary school students benefited least from the interventions compared to primary school students and adults, laboratory studies showed larger effects than classroom studies, and interventions were more effective for retrospective confidence judgments than for judgments of learning. For educational practice, interventions targeting the whole task, metacognitive knowledge, and external standards are recommended, while reconsideration and possibly discontinuation of timing interventions is needed.

Keywords Monitoring accuracy · Calibration · Metacognitive monitoring · Problem solving

✉ Noortje Janssen
noortje.janssen@ru.nl

[1] Behavioural Science Institute, Radboud University, PO Box 9104, 6500 HE Nijmegen, The Netherlands

Published online: 22 August 2024

Introduction

Problem solving is an important part of daily educational practice. Students learn to solve application problems in math classes, conduct science experiments in chemistry and physics, and propose solutions for socio-scientific issues in cross-curricular projects. Problem solving is also required for students' future professional life. From electricians who need to fix faulty wiring to event managers who need to find a solution for a sick catering supplier. Learning how to engage in understanding a new problem and finding a solution is thus one of the key elements to prepare students for their future jobs and society in general (Greiff et al., 2014; OECD, 2014).

When trying to solve a problem, students apply their knowledge to work toward a certain goal—the problem solution. For example, students in elementary school learn to solve multi-digit addition problems. To do so, they need to understand what addition entails, apply a solution strategy such as compensation, and change it if necessary to reach a solution (cf. De Jong & Ferguson-Hessler, 1996; Schoenfeld, 1979). After having solved the problem, students should accurately judge to what extent they mastered the problem-solving skill. This so-called monitoring accuracy is critical to learning as it causally relates to students' decision to continue or abort practicing. This decision, in turn, affects the successfulness of their problem solving in future tasks (Başokçu & Güzel, 2022; Jacobse & Harskamp, 2012; Rinne & Mazzocco, 2014).

Up until recently, problem solving received hardly any attention in monitoring accuracy research. Instead, most studies addressed the accuracy of judgments about learning lists of items (e.g., see reviews by Rhodes, 2016; Rhodes & Tauber, 2011) or texts (summarized in meta-analyses by Prinz et al., 2020a, b). This text-based focus was expanded in the 2010s when researchers realized that interventions designed to improve monitoring accuracy in text learning cannot be directly applied to problem-solving tasks (Baars et al., 2013; Kostons et al., 2010). Problem-solving tasks differ from learning items or studying texts in that students not only need to retrieve or understand information but also need to apply this information in such a way that they correctly perform problem-solving steps to reach a solution. Furthermore, judging the complex process of applying knowledge during problem solving is difficult and studies have consistently reported that students' monitoring accuracy in problem solving is generally poor (Başokçu & Güzel, 2022; Dentakos et al., 2019; García et al., 2016; Jacobse & Harskamp, 2012).

Bearing this theoretical and empirical evidence in mind, several researchers developed support to improve students' knowledge about their problem-solving skills, for example, by asking them to generate explanations (Pilegard & Fiorella, 2016) or answer metacognitive questions during the task (Gidalevich & Kramarski, 2019). Other studies sought to strengthen students' knowledge about the judgment itself by offering feedback on the accuracy of their performance and judgments (Kim, 2018; Oudman et al., 2022). The results of these studies were mixed: some interventions improved students' monitoring accuracy (Kim, 2018; Mihalca et al., 2015; Oudman et al., 2022), whereas others did not (Gidalevich & Kramarski, 2019; Kim, 2018; Pilegard & Fiorella, 2016).

This seemingly inconsistent evidence raises questions regarding the true effect of monitoring accuracy interventions in problem solving, and the factors that contribute to their effectiveness. Both questions were addressed in the present meta-analysis, which classified different types of interventions in terms of how they target students' judgments and investigated their relative effectiveness to improve monitoring accuracy in problem-solving tasks.

Theoretical Foundations of Monitoring Accuracy

Interventions aiming to improve monitoring accuracy are rooted in theories of metacognition. The seminal article by developmental psychologist Flavell (1979) defined metacognition as "knowledge and cognition about cognitive phenomena" (p. 906). This definition was further elaborated in the metamemory model by Nelson and Narens (1990), which specified the relations between cognitive processes, monitoring, and subsequent control in the learning context. This model consisted of two levels: the object level—which is about cognitive processes, and the meta level—which concerns metacognitive thinking and judgment processes. The meta level is informed by the object level through monitoring. Judgments on the meta level determine which control decisions are made on the object level, such as initiating changes in the learning process. So, according to this model, there is a causal connection between monitoring and control such that the accuracy of monitoring influences the quality of regulative actions undertaken by a student.

Monitoring and control processes play a central role in self-regulated learning (SRL) theories. These theories holistically describe how successful students self-regulate their learning: they set learning goals which they attempt to reach by monitoring and regulating their cognition, motivation, and behavior within the learning environment (Pintrich, 2000). To illustrate, the Zimmerman (2002) model posits that these processes take place in the phases of forethought (when self-assessment, goal setting, and planning processes occur), performance and volitional control (when students monitor and regulate learning strategies), and self-reflection (when students judge and attribute learning outcomes and set new plans for future learning). Other SRL models include similar phases with minor modifications depending on the researcher's perspective. One SRL-model that explicitly considers cognitive monitoring is Efklides' (2011) metacognitive and affective model of self-regulated learning (the MASRL model), which emphasizes the motivational components of metacognition and describes how motivational and affective learner characteristics interact with monitoring and control processes. Winne and Hadwin's (1998) conditions, operations, products, evaluations, and standards (COPES) model includes factors that influence cognitive monitoring and control processes during learning, such as task conditions and standards learners use to monitor their progress. Finally, the two-process model by Gutierrez et al. (2016) focuses on task-level monitoring by asking students to give judgments after a task. They include monitoring accuracy in their model and add monitoring error as a different but inversely related process.

The general assumption in the above-mentioned models is that the causal relation between monitoring and control extends to learning outcomes. Empirical research confirmed this assumption by showing that the accuracy of monitoring predicts control decisions and that both determine the extent to which the student masters the learning content (e.g., Thiede et al., 2003, 2012). The extent to which monitoring accuracy can be influenced is thus important to the success of the learning process.

To find out how students' monitoring accuracy can be improved, researchers have investigated what underlies students' judgment processes. This metamemory research typically instructed students to learn lists of items, such as words or word pairs, and asked them to judge how confident they would be in their performance on a later test (Rhodes, 2019). Using experimental designs, these studies found that students do not have direct access to their memory traces and therefore use information sources within the environment as cues to monitor their learning (Koriat, 1997), such as concreteness (Tauber & Rhodes, 2012, exp 4) or fluency of the to-be-learned item (Benjamin et al., 1998). The cue-utilization theory further holds that the use of cues does not necessarily result in correct judgments; only cues that are consistent with factors that affect performance would lead to improved monitoring accuracy (De Bruin & Van Merriënboer, 2017; Koriat, 1997). Thus, monitoring accuracy can be improved by interventions that target students' use of such appropriate cues.

Intervention Type as a Potential Moderator of Monitoring Accuracy

Insights from research on metamemory and text learning were taken as starting point for classifying interventions aiming to improve monitoring accuracy in problem-solving tasks. A broad distinction was made between interventions that focus on the learning task and interventions that directly address metacognitive judgment (Dunlosky & Lipko, 2007). Interventions addressing the learning task aim to focus students' attention on the cues within the learning content that are most informative for judging the extent to which this content is mastered. These interventions typically target students' use of cues during problem solving. If students use these cues effectively, their monitoring will be more accurate (Prinz et al., 2020b; Rawson et al., 2000). These types of interventions thus have an indirect impact on monitoring accuracy. Interventions addressing metacognitive judgment, by contrast, directly target monitoring accuracy by asking students to consider the process by which they determine how well they have mastered the learning content, that is, their metacognitive judgment after completing problem-solving tasks. If students carefully consider the correctness of their metacognitive judgment process, their monitoring accuracy is assumed to improve (Dunlosky & Lipko, 2007).

Interventions Addressing the Learning Task

As learning task interventions focus on cues in the learning content that are most informative for monitoring accuracy, these interventions depend on what students need to learn. In metamemory research students have to recall certain items and therefore need to use retrieval-based cues for their monitoring. Examples of effective interventions are practice and self-testing, which stimulate students to directly retrieve the to-be-recalled items (Rhodes, 2016). Effective interventions in the realm of text learning try to help students reach a deep understanding of the meaning of the text (Prinz et al., 2020b). For example, students are asked to create concept maps during reading to encode the information (Redford et al., 2012), or generate keywords or diagrams some time after reading the text so as to also retrieve the information (De Bruin et al., 2011; Thiede et al., 2005; Van Loon et al., 2014). Such *whole task interventions* thus focus on helping students understand what they (do not) know about the learning content.

In problem-solving tasks, students not only have to understand information but also have to apply it to a problem situation, which requires both *procedural knowledge* and *metacognitive knowledge* (Braithwaite & Sprague, 2021; Schoenfeld, 1979). Procedural knowledge concerns the content-related strategies to solve problems. Using the math example from the introduction, the compensation strategy requires students to perform three steps: (1) round each number to the nearest 10, (2) add them, and (3) subtract or add the difference to the original number. Metacognitive knowledge concerns thinking about the selection and application of problem-solving strategies. In the math example, students could opt for alternative strategies or could make mistakes executing the steps of the compensation strategy, for example, by omitting a step or by confusing addition and subtraction in step 3. Monitoring the applied strategies and changing them when mistakes are made is vital for efficient and effective problem solving (Schoenfeld, 1979, 2015). Attention for procedural and metacognitive knowledge seems therefore fruitful for interventions aiming to improve monitoring accuracy in problem solving.

Thus, in problem-solving tasks, interventions could take a whole task approach, similar to metamemory and text learning research. Additionally, interventions might address procedural knowledge to help students understand what they (do not) know about the content-related steps for problem solving, or might stimulate students to apply metacognitive strategies to help them consider how well they solved the problem.

Interventions Addressing Metacognitive Judgment

Interest in interventions aimed at metacognitive judgment has been instigated by Dunlosky and Lipko (2007), who found that even though interventions addressing the learning task improved monitoring accuracy in text learning, the accuracy scores remained rather low. They reasoned that interventions that directly target metacognitive judgment might better support students in reaching a correct judgment. With this reasoning in mind, three types of interventions can be distinguished: interventions aimed at the *timing* of judgments, availability of an *external standard*, and *monitoring training*.

Interventions that focus on the *timing* of judgments ask students to offer their judgment after a delay or ask them to more often judge their confidence during a learning task. Delayed judgments have mainly been applied in metamemory research. The idea behind this intervention is that the delay indirectly stimulates students to retrieve the information from memory, which consequently results in higher monitoring accuracy (Rhodes & Tauber, 2011). Asking students to more often judge their confidence has been investigated in early attempts to improve monitoring accuracy in text learning research. These judgments were elicited during a practice phase, while monitoring accuracy during the test phase was calculated to test their effectiveness (Bol et al., 2005; Nietfeld et al., 2006; Rawson & Dunlosky, 2007). Results were mixed and closer inspection of the studies revealed that monitoring accuracy improved only when students received feedback on their judgments during practice.

The reason that feedback improved monitoring accuracy when increasing judgments might be that these studies offered students an *external standard*—a point of reference to determine the extent to which their judgments were accurate. Students generally use internal standards for performance (i.e., what they think is a correct answer) when evaluating their work (Butler & Winne, 1995). However, especially for novice students such standards are generally of low quality which consequently results in inaccurate monitoring (Dunning et al., 2003). Offering external standards might alleviate the adverse effects of low-quality internal standards. Indeed, several studies have shown that students who received performance feedback——i.e., the correct answer (Lipko et al., 2009; Nederhand et al., 2019), or calibration feedback——i.e., the accuracy of their judgment (Callender et al., 2016, experiment 2; Geurten & Meulemans, 2017, experiment 1; Miller & Geraci, 2011, experiment 2) after practice problems were more accurate in their judgments on a later test than students who did not receive such feedback.

Finally, interventions that include *monitoring training* have not been applied in research on monitoring accuracy in metamemory and text learning, but might be especially effective for problem-solving tasks. As students need to engage in several problem-solving steps to reach a solution, they need to take each of these steps into account when judging the correctness of their problem solution. Research into students' performance during problem solving has shown that modeling examples effectively support them in systematically reflecting on problem-solving steps and improving their problem-solving procedures (e.g., Atkinson et al., 2000; van Gog & Rummel, 2010). Such interventions might therefore similarly support students in improving monitoring accuracy.

Previous Meta-Analyses on Interventions for Monitoring Accuracy

Previous meta-analyses on interventions aimed at improving monitoring accuracy examined the relative effectiveness of a certain type of intervention or focused on a specific type of task. Rhodes and Tauber (2011) did both by comparing delayed and immediate judgments during recall tasks. The effect of delayed judgments was robust ($g = 0.93$), indicating that delaying judgments improves students' monitoring accuracy.

In a meta-analysis on text learning tasks, the positive effect of delayed judgments could not be replicated (Prinz et al., 2020a). This might be because a delay mainly targets students' recall, but not their understanding of the information. In a follow-up meta-analysis, Prinz et al. (2020b) discovered that interventions that took a whole task approach by targeting text understanding improved monitoring accuracy by approximately half a standard deviation ($g = 0.46$).

Two recent meta-analyses did not specify the type of learning tasks. Gutierrez de Blume (2022) included studies incorporating a certain type of intervention, namely, learning strategy interventions. He identified a moderate effect ($g = 0.57$),[1] and found that deep strategy interventions resulted in higher monitoring accuracy than superficial strategies. León et al. (2023) investigated the effects of several types of interventions. However, they did not include monitoring accuracy, but focused on the closely related field of self-assessment accuracy—that is, the accuracy of students' assessment of "the quality of their own learning process and products" (Panadero et al., 2016, p. 804). They found a small effect ($g = 0.21$) with the strongest effects for external standards interventions.

These meta-analyses indicate that instructional interventions can improve monitoring accuracy during recall tasks and text learning tasks. However, it is still unclear whether and to what extent these benefits apply to problem-solving tasks. Equally unclear is whether and how this effectiveness depends on the design characteristics of the interventions used. Our meta-analysis extends these previous works by focusing on problem-solving tasks and including intervention types that target either the learning task or metacognitive judgment. Three intervention types addressed the learning task: (1) whole-task interventions that help students understand what they know about the problem, (2) interventions aimed at students' procedural knowledge, and (3) interventions aimed at students' metacognitive knowledge. Additionally, three intervention types addressed metacognitive judgment: (1) timing of the judgment, (2) providing external standards, and (3) training students to monitor their performance. Definitions and examples of these six intervention types are presented in Table 2.

Research Questions and Hypotheses

This meta-analysis set out to determine the overall effectiveness of monitoring accuracy interventions during problem solving (question 1) and how this effectiveness is moderated by intervention type (question 2). Based on the above literature review, we expected that monitoring accuracy interventions would have an overall positive effect (hypothesis 1), and that the magnitude of this effect would differ depending on type of intervention used (hypothesis 2). Additionally, we explored to what extent the six types of interventions described above differed in effectiveness, and how the overall mean effect size was moderated by the substantive characteristics *school level*, *domain*, and *judgment type*, and the methodological study features *research design* and *setting*.

[1] The original effect size was $g = -0.57$. In line with Rhodes and Tauber (2011), Prinz et al. (2020a, b), and to avoid confusion, we mirrored this effect size.

Method

The research questions, procedures, and analytic plan for this meta-analysis were preregistered and data were shared on the Open Science Framework (OSF; https://osf.io/u52w9/).

Literature Search

The literature search was inspired by Alexander's (2020) guidelines to systematically find potentially relevant research articles. The timeframe was restricted to the years 2002–2022 to include recent research on monitoring accuracy in problem-solving tasks. The electronic databases of Web of Science and ERIC were used for our main search. Web of Science was selected because it offers a wide access to research articles in the educational and psychology field. ERIC was selected because of its focus on educational research and because it includes research reports and dissertations by universities or non-profit organizations.

In Web of Science, we restricted our search to the abstract field and the categories Education Educational Research, Education Scientific Disciplines, Psychology, Psychology applied, and Psychology Educational. The database was searched using the following query: ("monitoring accuracy" OR "calibration" OR "metacognitive monitoring") AND ("effect*" OR "enhanc*" OR "improv*" OR "increas*"). As ERIC neither supports Boolean searches with more than six keywords nor truncations (Gusenbauer & Haddaway, 2020), we conducted separate searches for each possible query and replaced terms with truncations by the most likely term for an experimental research context——namely, effect, enhance, improve, and increase.

To extend the search results, Alexander (2020) recommends three actions: researcher checking, referential backtracking, and journal scouring. Researcher checking is the examination of the publication records of authors frequently appearing in the search results. As the first authors in our set differed widely, we performed this step by investigating all first authors' publication records in Web of Science. We additionally perused the publication records of two co-authors with more than five papers (Van Gog, $k=11$, and Paas, $k=9$). Referential backtracking involves examining references within important documents, such as related reviews. We included the articles that were meta-analyzed by Gutierrez de Blume (2022) in our pool of records for screening. Finally, journal scouring refers to hand searching journals that appear regularly in the search results. In our final set of records, not one journal stood out, so we skipped this action. The literature search was performed in January 2023 and the initial set of search outcomes contained 1945 records.

Study Selection

Figure 1 shows the steps taken to include the relevant studies for this meta-analysis. As a first step, we removed duplicates, which resulted in a total of 1736 unique records. Next, titles and abstracts were screened to determine whether the records

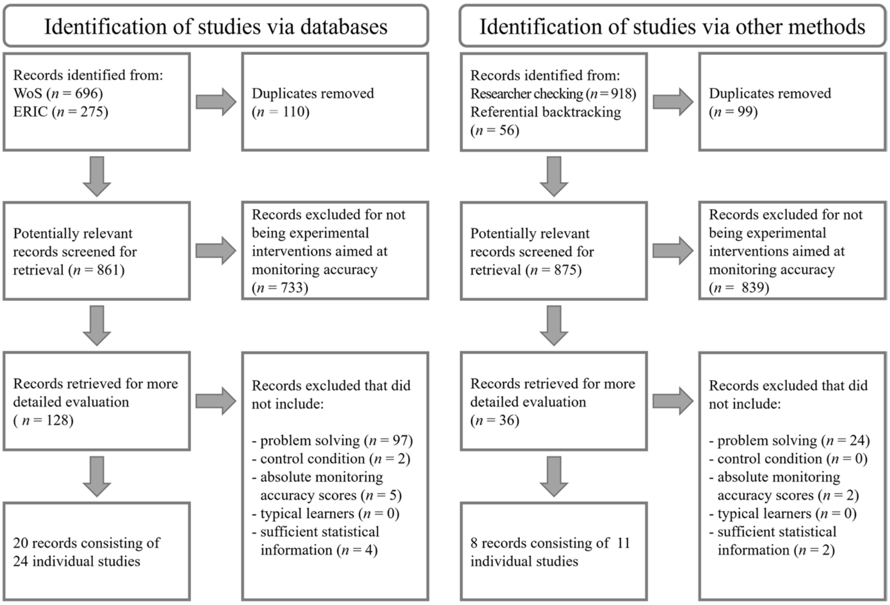
Identification of studies via databases
Identification of studies via other methods
Records identified from: WoS ($n = 696$) ERIC ($n = 275$)
Duplicates removed ($n = 110$)
Records identified from: Researcher checking ($n = 918$) Referential backtracking ($n = 56$)
Duplicates removed ($n = 99$)
Potentially relevant records screened for retrieval ($n = 861$)
Records excluded for not being experimental interventions aimed at monitoring accuracy ($n = 733$)
Potentially relevant records screened for retrieval ($n = 875$)
Records excluded for not being experimental interventions aimed at monitoring accuracy ($n = 839$)
Records retrieved for more detailed evaluation ($n = 128$)
Records excluded that did not include:
- problem solving ($n = 97$)
- control condition ($n = 2$)
- absolute monitoring accuracy scores ($n = 5$)
- typical learners ($n = 0$)
- sufficient statistical information ($n = 4$)
Records retrieved for more detailed evaluation ($n = 36$)
Records excluded that did not include:
- problem solving ($n = 24$)
- control condition ($n = 0$)
- absolute monitoring accuracy scores ($n = 2$)
- typical learners ($n = 0$)
- sufficient statistical information ($n = 2$)
20 records consisting of 24 individual studies
8 records consisting of 11 individual studies

Fig. 1 PRISMA flow chart of the search and selection of studies

administered a quantitative (quasi-)experimental between-subject design aimed at improving the participants' own monitoring accuracy. The 164 records that met this criterium were retrieved for more detailed information. Examples of excluded records were studies about calibration in other research fields (Goodney & Silverstein, 2013; Gorrini et al., 2018), theoretical or correlational studies (Hadwin & Webster, 2013; Hattie, 2013), studies that administered a within-subject design (Oudman et al., 2022), and studies of teachers' monitoring accuracy (Van De Pol et al., 2021). The full texts of retrieved records were screened for detailed evaluation of the tasks participants had to perform and the measures used. Records included satisfied the following criteria:

- Corresponded with our definition of problem-solving tasks. That is, tasks in which participants had to apply their knowledge to reach a certain goal—the problem solution. Included studies explicitly mentioned problem-solving tasks (e.g., Baars et al. 2014a; Mihalca et al., 2015); used generally known problem-solving tasks, such as the Raven's test task (Mitchum & Kelley, 2010) or the Latin square task (Double & Birney, 2018); or administered tasks that did not explicitly mention problem solving but required problem-solving steps to reach a solution, such as a location in navigational map reading (Kok et al., 2022) or an answer to a research question in inquiry tasks (Kant et al., 2017; Morrison et al., 2015). Excluded studies were mostly about text learning (e.g., Huff & Nietfeld, 2009; Kimball et al., 2012; Thiede et al., 2011) or lacked a clear description of the learning tasks (e.g., Bingham et al., 2021; Bol et al., 2005; Morphew, 2021; Osterhage, 2021).

- Contained at least one intervention condition that aimed to improve students' monitoring accuracy and a control condition that was either untreated or included an intervention that did not aim to improve monitoring accuracy.
- Reported absolute monitoring accuracy scores (see Schraw, 2009). In case bias scores were reported, the corresponding author was asked to e-mail us a conversion into absolute accuracy scores.
- Addressed typical students in formal education and adults.
- Contained sufficient statistical information to calculate effect sizes. If this was not the case, authors were e-mailed to request the missing information.

The final set of 28 records consisted of 35 independent studies; their characteristics are presented in Table 1.

Outcome Measure

Monitoring accuracy is defined as the deviation of students' judgments from their actual performance. The most common monitoring accuracy measures in problem-solving research are bias and absolute accuracy. Bias is calculated by subtracting actual performance scores from students' judgment scores. Although this metric offers valuable information about the direction of students' judgment, it does not reflect the magnitude of judgment error. The absolute accuracy measure does because it represents the absolute deviance between students' judgments and their performance (Schraw, 2009). This meta-analysis therefore used the absolute accuracy scores as outcome measure, which was calculated in the primary studies either by squaring or recoding the deviation between students' judgments and performance.

Moderator Variables

Six moderators were included in this meta-analysis. The first moderator, *intervention type*, defined the instructional interventions taken to improve students' monitoring accuracy. Table 2 offers an overview of the types of interventions included in this meta-analysis. Three intervention types focus on the learning task: whole task, procedural knowledge, and metacognitive knowledge. Intervention types targeting metacognitive judgment manipulate the timing of the judgment, offer external standard support, or give monitoring training. To determine interrater agreement on the classification of the interventions used in the included studies, two independent raters coded the interventions of all 35 studies and discussed differences until agreement was reached. Disagreements were randomly distributed across codes and the Cohen's κ of 0.79 showed that agreement between raters was substantial (cf. Landis & Koch, 1977).

Substantive characteristics encompassed school level, domain, and judgment type. *School level* was used to examine whether and how students' age moderated the findings. We distinguished between three age groups, based on the most common school systems: elementary school children (ages 5–12), secondary school adolescents (ages 13–17), and adults (age $\geq$ 18). Based on the tasks used in the

Table 1 Studies included in this meta-analysis

Author(s)	n_{control}	$n_{\text{treatment}}$	g	Intervention type	School level	Domain	Judgment[a]	Design[b]	Setting
Baars et al. (2018a, exp 1)	38	40	0.14	Procedural knowledge	Secondary	Science	JOL	Exp	Classroom
Baars et al. (2018a, exp 2)	27	18	0.05	Procedural knowledge	Secondary	Science	JOL	Exp	Classroom
Baars et al. (2014a)	53	82	0.11	Whole task	Elementary	Math	JOL	Exp	Classroom
Baars et al. (2017)	57	78	0.43	Whole task	Secondary	Science	JOL	Exp	Classroom
Baars et al. (2018b)	34	40	−0.34	Timing	Elementary	Math	JOL	Exp	Classroom
Baars et al. (2014b, exp 1)	21	23	0.27	Training	Secondary	Science	JOL	Exp	Classroom
Baars et al. (2014b, exp 2)	67	66	0.14	External standard	Secondary	Science	JOL	Exp	Classroom
Baars et al. (2013)	32	29	−0.16	Whole task	Secondary	Science	JOL	Exp	Classroom
Digiacomo and Chen (2016)	15	15	0.43	Mixed	Elementary	Math	RCJ	Exp	Classroom
Double and Birney (2018)	46	43	−0.45	Timing	Adults	Reasoning	RCJ	Exp	Classroom
Ford (2018)	15	18	0.39	Metacognitive knowledge	Secondary	Math	RCJ	Q exp	Classroom
Fyfe et al. (2022)	66	70	0.60	Metacognitive knowledge	Elementary	Math	RCJ	Exp	Classroom
Kant et al. (2017)	19	18	0.19	Whole task	Secondary	Science	JOL	Exp	Classroom
Kok et al. (2022)	30	34	0.16	Whole task	Adults	Other	RCJ	Exp	Laboratory
Kostons et al. (2012, exp 1)	40	40	0.12	Training	Secondary	Science	RCJ	Exp	Classroom
Kostons et al. (2012, exp 2)	33	57	0.35	Training	Secondary	Science	RCJ	Exp	Classroom
Kuhn et al. (2022)	35	34	−0.37	External standard	Adults	Other	RCJ	Exp	Classroom
Labuhn et al. (2010)	30	60	0.69	External standard	Elementary	Math	RCJ	Exp	Classroom
Mihalca and Mengelkamp (2020)	82	83	0.78	Whole task	Adults	Science	RCJ	Exp	Laboratory
Mitchum and Kelley (2010, exp 2)	29	28	0.60	Whole task	Adults	Reasoning	RCJ	Exp	Laboratory
Morrison et al. (2015)	28	57	0.56	Whole task	Adults	Science	RCJ	Exp	Laboratory
Nietfeld and Schraw (2002, exp 1)	31	28	0.55	Whole task	Adults	Math	RCJ	Q exp	Classroom
Nietfeld and Schraw (2002, exp 2)	26	32	0.61	Procedural knowledge	Adults	Math	RCJ	Exp	Classroom
Raaijmakers et al. (2018)	26	52	−0.12	Training	Secondary	Math	RCJ	Exp	Classroom

Table 1 (continued)

Author(s)	n_{control}	$n_{\text{treatment}}$	g	Intervention type	School level	Domain	Judgment[a]	Design[b]	Setting
Raaijmakers et al. (2019, exp 1)	37	72	−0.41	External standard	Secondary	Science	RCJ	Exp	Classroom
Raaijmakers et al. (2019, exp 2)	36	74	0.08	External standard	Secondary	Science	RCJ	Exp	Classroom
Ramdass and Zimmerman (2008)	21	21	0.72	External standard	Elementary	Math	RCJ	Exp	Classroom
Saenz et al. (2019)	33	38	0.65	External standard	Adults	Reasoning	RCJ	Exp	Laboratory
Sieck and Arkes (2005, exp 1)	58	52	−0.02	External standard	Adults	Reasoning	RCJ	Exp	Not reported
Sieck and Arkes (2005, exp 2)	45	35	0.21	Metacognitive knowledge	Adults	Reasoning	RCJ	Exp	Not reported
Sieck and Arkes (2005, exp 3)	37	38	0.39	External standard	Adults	Reasoning	RCJ	Exp	Not reported
Testa et al. (2020)	201	182	0.02	Mixed	Secondary	Science	RCJ	Exp	Classroom
Van Loon and Roebers (2020)	35	70	0.91	External standard	Elementary	Reasoning	RCJ	Exp	Laboratory
Wang et al. (2021)	44	45	0.20	Mixed	Secondary	Math	RCJ	Q exp	Classroom
Wollenschlager et al. (2016)	40	80	0.36	External standard	Secondary	Science	JOL	Exp	Classroom

[a] *JOL*, judgment of learning; *RCJ*, retrospective confidence judgment

[b] *Exp*, experimental design; *Q exp*, quasi-experimental design

$N=3219$

Table 2 Classification of the interventions on monitoring accuracy in problem-solving tasks

Intervention type	Description	Examples
Learning task	Focus students' attention on the correct cues during problem solving	
Whole task	Help students understand what they (do not) know about the problem	Students are asked to perform practice problems (Baars et al., 2014a)
Procedural knowledge	Help students understand the content-related steps to solve a problem	Students receive a strategy training to solve probability problems (Nietfeld & Schraw, 2002)
Metacognitive knowledge	Help students perform the metacognitive steps to understand what they (do not) know to solve the problem	Students receive metacognitive questions during the task, such as "what steps do I need to take to get the right answer?" (Fyfe et al., 2022)
Metacognitive judgment	Ask students to consider their metacognitive judgment after problem solving	
Timing	Control *when* students make judgments	Students are asked to make judgments some time after the task (Baars et al., 2018b)
External standard	Give students a benchmark against which to assess their performance	Students receive the correct answer after making their judgment (Van Loon & Roebers, 2020)
Monitoring training	Train students to monitor their performance	Students watch video models who judge the problem-solving steps of a task they have completed (Kostons et al., 2012)
Mixed	(All possible combinations of the above intervention types)	Students are instructed to use Zimmerman's model during the task *and* receive the correct answer afterward (Digiacomo & Chen, 2016)

primary studies, the problem-solving *domain* was classified as either mathematics, science, or reasoning. Mathematics included problem-solving tasks in mathematics education. Science comprised tasks in the exact disciplines of physics, chemistry, and biology. Reasoning involved logical, inductive, and analogical reasoning tasks. Lastly, we used an "other" category that comprised domains not included above. The moderator *judgment type* indicated whether the monitoring accuracy score was based on retrospective confidence judgments (RCJs) or judgments of learning (JOLs). RCJs ask students to indicate their confidence in the correctness of their performance on a completed problem-solving task, whereas JOLs ask students to make a predictive judgment of how confident they are regarding their performance on a future problem-solving task. Note that this is a conceptual distinction: in both cases, monitoring accuracy is calculated as the difference between the students' judgment (either RCJ or JOL) and actual performance.

Methodological study features included design and setting. *Design* was classified as experimental or quasi-experimental designs. *Setting* involved the amount of experimental control and was classified as either classroom or laboratory experiments. Classroom experiments had low experimental control, for example, authentic classroom settings, but also cases where adults were performing the problem-solving tasks in an online setting. Laboratory experiments had high experimental control. Examples are experiments with individuals or small groups and quiet settings where students were individually tested in school. When it was unclear in which setting the study took place, it was coded as "not reported."

Effect Size Calculation

Hedges' *g* was used as effect size measure because it includes a correction for small sample sizes (Borenstein et al., 2009). Effect sizes were calculated based on the means and standard deviations of monitoring accuracy scores in the intervention condition and control condition. As more accurate monitoring implies smaller deviations, lower scores reflect higher accuracy. Such positive results are thus reflected in negative effect sizes. To avoid confusion, the effect sizes were mirrored so that a positive effect size reflects higher accuracy scores in the experimental conditions.

In studies that provided both pretest and posttest monitoring accuracy scores, the mean gains and pre- and posttest standard deviations were used to calculate effect sizes (Lipsey & Wilson, 2001). In studies that included multiple outcomes (e.g., multiple difficulty levels or timepoints), these outcome measures were collapsed into a composite effect size via the formulas by Borenstein et al. (2009). When studies only presented outcomes for subgroups within each condition, the subgroups' means and standard deviations were merged (Borenstein et al., 2009) and the effect size was calculated (Lipsey & Wilson, 2001).

In studies investigating multiple interventions, a composite effect size value was calculated if the interventions were of the same type (e.g., completion problems and practice problems are both whole-task interventions), following the same procedure as studies that only presented outcomes for subgroups. If different types of interventions were investigated, the intervention that matched the study's main purpose was selected. If this was not possible, one intervention was selected at random.

Data Analysis

Main analyses were conducted with Meta Essentials (Suurmond et al., 2017). We first analyzed the effect sizes for signs of publication bias through visual inspection of the funnel plot, Begg and Mazumdar's (1994) rank correlation test,[2] and Egger et al.'s (1997) regression test. Next, the overall mean effect size was calculated and a random effects model was applied to examine whether the mean effect differed significantly from zero (hypothesis 1). A Q-test was conducted to analyze whether intervention type moderated the findings (hypothesis 2); significant p-values indicate that the true effects vary between intervention types. Pairwise comparisons (Hedges & Pigott, 2004) were conducted to investigate differences in effect size between the six intervention types. Hochberg's (1988) step-up procedure was used to control the familywise type I error at $\alpha = 0.05$. Adjusted p-values were reported for $\alpha = 0.05$.

In a series of exploratory analyses, which were not included in our pre-registration, Q-tests were used to investigate whether school level, domain, judgment type, design, and setting moderated the findings. Pairwise comparisons and alpha-level corrections were conducted similarly to the moderator analysis of intervention type.

Results

Publication Bias

The funnel plot in Fig. 2 indicates low chance of publication bias as the studies were evenly distributed around the combined effect size. The follow-up tests statistically confirmed this symmetric distribution. The rank-correlation test showed a low and nonsignificant correlation coefficient, Kendall's $T = 0.08$, $z = 0.64$, $p = 0.523$, and in the regression test the intercept did not differ significantly from zero, $\beta = 0.90$, 95% CI [-1.29, 3.08], $p = 0.411$. This means that both tests did not show evidence of publication bias.

Overall Effect

The overall mean effect size (g) of the 35 included studies was 0.25, $SE = 0.06$, 95% CI [0.12, 0.37]. This result indicates that all interventions together had a small effect (Cohen, 1988) on monitoring accuracy. The magnitude of the overall effect size differed significantly from zero, $z = 4.09$, $p < 0.001$. The I^2 statistic showed that 63.62% of the variance in effect sizes reflects true score variation; the estimated variance of the true effect size (T^2) was 0.08.

[2] This test was conducted at a reviewer's request and was not part of our preregistration.

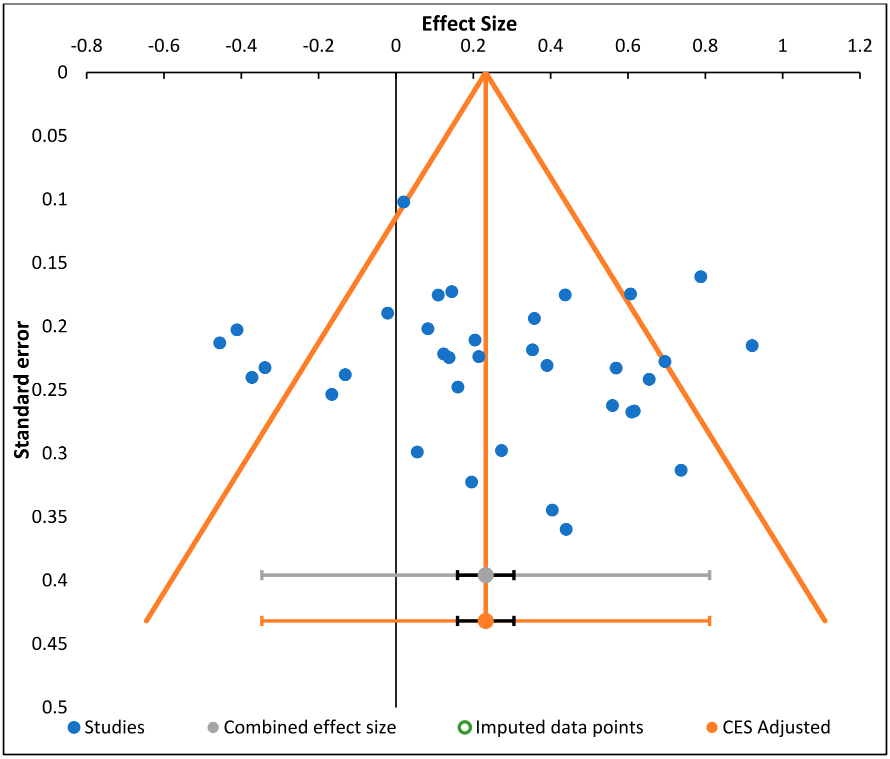

Fig. 2 Funnel plot of effect sizes

Moderator Analysis: Intervention Type

The heterogeneity of effect sizes was scrutinized through moderator analysis. As shown in Table 3, intervention type significantly moderated the findings. The 95% confidence intervals indicate that the intervention types whole task, metacognitive knowledge, and external standards all significantly improved students' monitoring accuracy. On the other hand, the timing interventions significantly decreased students' monitoring accuracy. The other interventions did not significantly impact monitoring accuracy. Pairwise comparisons of all interventions types—except the mixed interventions—showed that timing had a significantly lower effect size than all the other intervention types, with values ranging from $z = -4.56$, $p < 0.001$ for whole task to $z = -2.76$, $p = 0.017$ for training. All other intervention types did not significantly differ in their effectiveness, with p-values ranging from 0.237 to 0.968.

Table 3 Summary of effect sizes per moderator variable category

	k	*g*	95% CI	*Q*	df	*p*	I^2
Intervention type				23.86	6	0.002	74.85
Whole task	9	0.38	[0.18, 0.58]				
Metacognitive knowledge	3	0.45	[0.20, 0.70]				
Procedural knowledge	3	0.27	[−0.07, 0.60]				
Timing	2	−0.40	[−0.52, −0.29]				
External standard	11	0.28	[0.01, 0.54]				
Training	4	0.15	[−0.06, 0.36]				
Mixed	3	0.08	[−0.08, 0.24]				
School level				4.72	2	0.189	57.63
Elementary	7	0.45	[0.12, 0.77]				
Secondary	16	0.12	[0.01, 0.23]				
Adults	12	0.31	[0.07, 0.54]				
Domain				2.86	3	0.470	0.00
Math	11	0.34	[0.12, 0.55]				
Science	15	0.21	[0.05, 0.36]				
Reasoning	7	0.32	[−0.03, 0.67]				
Other	2	−0.11	[−0.64, 0.42]				
Judgment type				1.94	1	0.245	48.45
RCJ	25	0.30	[0.14, 0.45]				
JOL	10	0.15	[0.00, 0.29]				
Design				0.52	1	0.470	0.00
Experimental	32	0.24	[0.11, 0.37]				
Quasi-experimental	3	0.36	[0.13, 0.58]				
Setting				17.09	2	0.001	88.30
Classroom	26	0.16	[0.03, 0.29]				
Laboratory	6	0.65	[0.45, 0.85]				
Not reported	3	0.17	[−0.07, 0.41]				

Exploratory Moderator Analyses: Substantive and Methodological Study Features

As shown in Table 3, the interventions significantly and positively impacted monitoring accuracy in all study features, except the learning domains “reasoning” and “other,” and studies that did not report their setting. Results of the *Q*-tests show that setting was the only study feature that significantly moderated the findings. However, a nonsignificant *p*-value does not necessarily prove that the effect sizes are consistent; a lack of significance may be due to low power because of the limited number of studies (Borenstein et al., 2009). We therefore explored differences between moderator categories for each moderator variable by means of pairwise comparisons, with the exception of the moderators *domain* and *design* due to similarities in confidence intervals and the 0.00% score on I^2.

Regarding students' school level, interventions were generally less effective for secondary school students than for elementary school students, $z = -4.75$, $p < 0.001$, and adults, $z = -2.14$, $p = 0.049$, while differences between elementary school students and adults were not significant, $z = 1.28$, $p = 0.201$. Interventions were more effective for retrospective confidence judgments than judgments of learning, $z = 2.03$, $p = 0.042$, and studies conducted in laboratory settings had significantly higher effect sizes than classroom studies, $z = 4.98$, $p < 0.001$.

Discussion

This research built on prior meta-analyses by examining the effectiveness of interventions aimed at improving monitoring accuracy in the context of problem solving. In line with the first hypothesis, the interventions significantly impacted monitoring accuracy and there was little evidence of publication bias. However, with $g = 0.25$ this effect was small, especially compared to effects found in other meta-analyses on monitoring accuracy interventions (e.g., Gutierrez de Blume, 2022: $g = 0.57$; Prinz et al., 2020b: $g = 0.46$). Furthermore, not every intervention type positively and significantly affected monitoring accuracy. Interventions on the whole task, metacognitive knowledge, and external standards had a positive impact, procedural knowledge, mixed and training were not significant, and timing interventions negatively impacted monitoring accuracy. In line with hypothesis 2, moderator analysis showed that effects varied between intervention types. Timing interventions had a negative effect on monitoring accuracy and significantly differed from all other intervention types. Exploratory moderator analyses revealed study feature effects. School level, judgment type, and setting need consideration in future research.

The relatively small overall effect of the interventions on monitoring accuracy might be due to the large variation in true effect sizes. This is most apparent in the difference between timing and the other intervention types, although variations within and between other intervention types were considerable, too. Another reason for the small effect could be related to the complexity of the tasks in which the interventions were implemented. Previous meta-analyses mainly included recall and text learning tasks that do not require students to apply information to a certain situation. Problem solving does and therefore additionally takes up procedural and metacognitive knowledge (Braithwaite & Sprague, 2021; Schoenfeld, 1979). Combining and applying these knowledge components requires a considerable amount of students' working memory (Cornoldi et al., 2015; Sweller, 2023; Sweller et al., 1998) and might have reduced the impact of the interventions. Future research should investigate whether this is indeed the case and consider other possible factors that might have influenced the effect size of the interventions.

The reason that timing stood out compared to the other interventions might also be related to the complexity of problem-solving tasks. Previous meta-analyses already showed that students benefit from timing interventions in recall tasks (Rhodes & Tauber, 2011), but not in text learning (Prinz et al., 2020a). This meta-analysis adds that its effect even reverses in problem-solving tasks. The two studies that included this intervention type differed in how they changed the timing of judgments. Baars, Van Gog et al. (2018b) asked elementary school students to judge their math performance after a time delay, while Double and Birney (2018) asked adults to more often judge their confidence during the Latin square task. Authors of both studies reasoned that their intervention retained participants from using experience-based cues. Specifically, a time delay might have made it difficult to use the cues that were most salient during problem solving (Baars et al. 2018b), while increasing judgment frequency focused participants' attention on their existing confidence-related beliefs. The latter was confirmed by a follow-up analysis that showed that participants in the intervention condition relied more on judgments they made prospectively than the control participants (Double & Birney, 2018). It thus seems that, although studies are few, timing interventions do not to draw students' attention to their problem-solving experiences and therefore negatively impact monitoring accuracy.

Several intervention types had no significant impact on monitoring accuracy. The monitoring training interventions might have been ineffective because they focused on monitoring the problem-solving steps without offering an external standard against which students could evaluate their knowledge (Dunlosky & Lipko, 2007). Procedural knowledge interventions were included in three studies. Nietfeld and Schraw (2002, exp 2) offered university students a strategy training for probability problems and found this intervention to be effective, whereas Baars et al. (2018a, 2018b) found no effects of self-explaining the steps in solving biology problems in two studies with secondary school students. Perhaps the sole focus on explaining problem-solving steps might not sufficiently improve monitoring accuracy because students also needed to consider their conceptual understanding and metacognitive thinking (Braithwaite & Sprague, 2021; Schoenfeld, 1979). For mixed interventions it is difficult to draw conclusions, as these interventions differed in many ways.

Another reason why the above-mentioned interventions are seemingly ineffective for improving monitoring accuracy might be related to the students' school level. In all these intervention types, at least two-third of the studies were conducted in secondary education, and in line with the results of this moderator variable, these studies lowered the overall effect size of the intervention types. Given the limited number of studies on these intervention types, with the majority conducted in secondary schools, future research should further investigate the effectiveness of these intervention types in different contexts, starting with primary school students and adults.

The finding that interventions were generally less effective for secondary school students compared to primary school students and adults offers new insight in the effectiveness of monitoring accuracy interventions. Until now, only the meta-analysis by Gutierrez de Blume (2022) took age into account. He found that interventions were more effective for adults than for children. A reason for the lower effect size scores in secondary school in this meta-analysis might be related to students' development in motivational beliefs. Classical research on students' motivation shows a decline in school-related motivational beliefs during adolescence (e.g., Eccles et al., 1993; Wigfield et al., 1991) and recent research confirms that secondary school students with low motivation learn less from interventions aimed at improving monitoring accuracy than students with moderate or high motivation (Wijnia & Baars, 2021). Thus, the general decline in motivation during the secondary school years might have resulted in lower effects of monitoring accuracy interventions. Still, considering the limited number of studies and nonsignificant *Q*-statistics, future research should first replicate the present findings with a larger set of studies. As these are not available in the problem-solving field, the impact of school level should be tested in related fields that include interventions on monitoring accuracy such as text learning.

Although setting made a difference in that laboratory studies showed higher effects than classroom studies, it should be noted that interventions in classroom studies also had a significantly positive impact on monitoring accuracy. This is in line with the meta-analysis by Gutierrez de Blume (2022) and suggests that despite these more "messy" settings, such interventions can be recommended for use in authentic classrooms. Regarding judgment type, pairwise comparisons revealed that interventions had a greater effect on retrospective confidence judgments than judgments of learning. This might be because retrospective confidence judgments ask students to recall their performance on a task they just completed, while judgments of learning are prospective and therefore require students to predict an uncertain future (Dougherty et al., 2005). Future research should investigate how to optimize monitoring accuracy interventions in authentic classrooms for both retrospective and prospective judgments.

Limitations

With 35 included studies in the meta-analysis, the number of studies that aim to improve monitoring accuracy in problem-solving is relatively low. This number could have been higher when all retrieved studies had reported sufficient information to calculate effect sizes. Ten studies lacked this statistical information, and authors of only four of these studies sent us the requested information. The others did not respond or did not have the required data. Not only for future authors but also reviewers and editors, it is thus important to be cognizant of whether sufficient statistical information is available when studies are published so that they can be included in the growing body of meta-analyses in educational psychology.

The low number of studies within some intervention types has consequences for the strength of the conclusions that can be drawn. A good example is that timing included only two studies. Although it was clearly found that timing interventions did not improve monitoring accuracy, with only two studies there is a higher chance that other study-related factors could have influenced the results. A related consequence of the low number of studies was that the *Q*-statistics were not significant for the moderators school level and judgment type while they were for the pairwise comparisons. As the pairwise comparisons and *Q*-tests are different tests, it is reasonable to anticipate different results, especially due to the *Q*-test's sensitivity to low power (Borenstein et al., 2009). To ensure no false positives were reported, Hochberg's alpha correction procedure was used. However, as each study has unique contextual features, such as the sample, materials, and procedure, caution is needed when generalizing these results to educational practice. Nevertheless, these moderator analyses offer sufficient starting points for future research on interventions for monitoring accuracy.

Implications

This meta-analysis was conducted in the context of theories that pose that SRL is beneficial to students' long-term learning and development (Dignath et al., 2023). Early SRL interventions aimed to improve students' regulation during all phases of SRL models (e.g., De Boer et al., 2014; Dignath & Büttner, 2008; Dignath et al., 2008), while more recently, meta-analyses addressed interventions on *monitoring* and their effectiveness for students' learning outcomes (Dignath et al., 2023; Guo, 2022a, b). The current meta-analysis is even more specific, as it contributes to SRL by investigating the *accuracy* of students' monitoring and demonstrating that this accuracy can be improved.

Specifically, the results of this meta-analysis show that interventions addressing the whole task, metacognitive knowledge, and external standards positively impact monitoring accuracy. These results are largely consistent with previous meta-analyses (Gutierrez de Blume, 2022; León et al., 2023; Prinz et al., 2020b) and suggest that the effectiveness of these three interventions generalizes across tasks and domains. To better understand the mechanisms underlying these successful interventions, future research should investigate whether different design configurations lead to differential effectiveness. For example, a good starting point might be to investigate whether an intervention type should offer support on all steps of the problem-solving process or whether it should only support the final step—i.e., the solution.

Theoretical implications relate to the incorporation and support of monitoring accuracy in models of (meta)memory and SRL. Our results indicate that monitoring accuracy can be improved and that successful interventions give students insight into their cognitive and metacognitive processes during problem solving and support the monitoring of their performance. Monitoring accuracy can thus be improved by providing support on each component of Nelson and Narens' (1990) metamemory model, which suggests that this descriptive model can be used as a normative framework for research and development of successful instructional interventions. Our results substantiate the pivotal role of standards in the COPES model (Winne & Hadwin, 1998) by showing that monitoring accuracy improves if students evaluate their performance against an offered, and hence appropriate benchmark. Finally, implications for the monitoring model of Gutierez et al. (2016) are less straightforward. In line with their conclusion of a domain general monitoring model, our moderator analysis found that the effectiveness of interventions was independent of the task domain. However, we could not find direct evidence for the domain-general notion that older individuals would be more proficient in monitoring, as adults benefitted as much from the interventions as children did. These results suggest that monitoring accuracy with additional support differs from unsupported monitoring and that successful interventions might compensate for age-related differences in monitoring accuracy. Future research should investigate both suggestions.

For educators, this meta-analysis provides evidence to suggest the use of interventions to improve students' monitoring accuracy in problem-solving tasks. Specifically, interventions addressing the whole task, metacognitive knowledge, and external standards are recommended for problem solving given their positive impact on monitoring accuracy. On the other hand, as timing adversely affected students' monitoring accuracy, educators should refrain from implementing this intervention until further research is conducted to ascertain its effectiveness.

To conclude, the current meta-analysis contributes to the growing body of evidence addressing monitoring accuracy in the context of SRL. Several factors not included in this meta-analysis might be of interest to future meta-analyses on monitoring accuracy, for example, the effort students (are willing to) put into their learning (De Bruin et al., 2020; Van Gog et al., 2020) or task characteristics (cf. Prinz et al., 2020a), such as the amount of problem-solving steps or problem-solving performance during practice tasks. The studies that include such factors are few but promising (Baars et al., 2014b; Kostons et al., 2012; Mihalca et al., 2017). When more studies are available, these could give a more comprehensive understanding of how to best support students in improving their monitoring of problem solving, their regulatory decisions, and consequently the overall efficiency of learning problem-solving skills.

Acknowledgements We are grateful to Mare van Hooijdonk, who scored part of the data to determine interrater reliability. We also thank Martine Baars for her feedback on a preliminary version of this manuscript.

Author Contribution The first author came up with the research idea and drafted the theoretical framework, design, analyses, results and discussion. The second author contributed his experience to conduct the meta-analysis procedures, commented on and edited previous versions of the manuscript, and read and approved the final manuscript.

Data Availability The data for this meta-analysis were shared on the Open Science Framework (OSF; https://osf.io/u52w9/).

Declarations

Preregistration The research questions, procedures, and analytic plan for this meta-analysis are preregistered on the Open Science Framework (OSF; https://osf.io/xsq78.).

Ethics Approval This is a meta-analysis for which no ethical approval is required.

Conflict of Interest The authors declare no competing interests.

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research, 90*(1), 6–23. https://doi.org/10.3102/0034654319854352

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Baars, M., Visser, S., Gog, T. V., Bruin, A. D., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology, 37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Baars, M., Leopold, C., & Paas, F. (2018a). Self-explaining steps in problem-solving tasks to improve self-regulation in secondary education. *Journal of Educational Psychology, 110*(4), 578–595. https://doi.org/10.1037/edu0000223

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2018b). Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks. *Studies in Educational Evaluation, 58*, 51–59. https://doi.org/10.1016/j.stueduc.2018.05.010

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2014a). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382–391. https://doi.org/10.1002/acp.3008

Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2014b). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004

Başokçu, T. O., & Güzel, M. A. (2022). Beyond counting the correct responses: Metacognitive monitoring and score estimations in mathematics. *Psychology in the Schools*. https://doi.org/10.1002/pits.22665

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088–1101. https://doi.org/10.2307/2533446

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Bingham, B. E., Coulter, C., Cottenie, K., & Jacobs, S. R. (2021). A metacognitive instructional guide to support effective studying strategies. *The Canadian Journal for the Scholarship of Teaching and Learning*, *12*(1), Art. 5. https://doi.org/10.5206/cjsotlrcacea.2021.1.8318

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290. https://doi.org/10.3200/jexe.73.4.269-290

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. *John Wiley & Sons, Ltd.* https://doi.org/10.1002/9780470743386

Braithwaite, D. W., & Sprague, L. (2021). Conceptual knowledge, procedural knowledge, and metacognition in routine and nonroutine problem solving. *Cognitive Science*, *45*(10). https://doi.org/10.1111/cogs.13048

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cornoldi, C., Carretti, B., Drusi, S., & Tencati, C. (2015). Improving problem solving in primary school students: The effect of a training programme focusing on metacognition and working memory. *British Journal of Educational Psychology, 85*(3), 424–439. https://doi.org/10.1111/bjep.12083

De Boer, H., Donker, A. S., & Van Der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509–545. https://doi.org/10.3102/0034654314540006

De Bruin, A. B. H., Roelle, J., Carpenter, S. K., & Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review, 32*(4), 903–915. https://doi.org/10.1007/s10648-020-09576-4

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 Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1–9. https://doi.org/10.1016/j.learninstruc.2017.06.001

De Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105–113. https://doi.org/10.1207/s15326985ep3102_2

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

Digiacomo, G., & Chen, P. P. (2016). Enhancing self-regulatory skills through an intervention embedded in a middle school mathematics curriculum. *Psychology in the Schools, 53*(6), 601–616. https://doi.org/10.1002/pits.21929

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*(3), 231–264. https://doi.org/10.1007/s11409-008-9029-x

Dignath, C., Büttner, G., & Langfeldt, H. (2008). How can primary school students acquire self-regulated learning most efficiently? A meta-analysis on interventions that aim at fostering self-regulation. *Educational Research Review, 3*(2), 101–129.

Dignath, C., Van Ewijk, R., Perels, F., & Fabriz, S. (2023). Let learners monitor the learning content and their learning behavior! A meta-analysis on the effectiveness of tools to foster monitoring. *Educational Psychology Review*, *35*(2). https://doi.org/10.1007/s10648-023-09718-4

Double, K. S., & Birney, D. P. (2018). Reactivity to confidence ratings in older individuals performing the latin square task. *Metacognition and Learning, 13*(3), 309–326. https://doi.org/10.1007/s11409-018-9186-5

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*(6), 1096–1115. https://doi.org/10.3758/bf03193216

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Iver, D. M., & Feldlaufer, H. (1993). Negative effects of traditional middle schools on students' motivation. *The Elementary School Journal*, *93*(5), 553–574. http://www.jstor.org/stable/1001828

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

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

*Ford, D. J. (2018). *The effects of metacognitive training on algebra students' calibration accuracy, achievement, and mathematical literacy* [Unpublished doctoral dissertation]. Old Dominion University.

Fyfe, E. R., Byers, C., & Nelson, L. J. (2022). The benefits of a metacognitive lesson on children's understanding of mathematical equivalence, arithmetic, and place value. *Journal of Educational Psychology, 114*(6), 1292–1306. https://doi.org/10.1037/edu0000715

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*(2), 139–170. https://doi.org/10.1007/s11409-015-9139-1

Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive judgments: A heuristic account. *Journal of Cognitive Psychology, 29*(2), 184–201. https://doi.org/10.1080/20445911.2016.1229669

Gidalevich, S., & Kramarski, B. (2019). The value of fixed versus faded self-regulatory scaffolds on fourth graders' mathematical problem solving. *Instructional Science, 47*(1), 39–68. https://doi.org/10.1007/s11251-018-9475-z

Goodney, D. E., & Silverstein, T. P. (2013). Using the tyrosinase-based biosensor to determine the concentration of phenolics in wine. *Journal of Chemical Education, 90*(12), 1710–1712. https://doi.org/10.1021/ed300495a

Gorrini, A., Crociani, L., Vizzari, G., & Bandini, S. (2018). Observation results on pedestrian-vehicle interactions at non-signalized intersections towards simulation. *Transportation Research Part F-Traffic Psychology and Behaviour, 59*, 269–285. https://doi.org/10.1016/j.trf.2018.09.016

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13*, 74–83. https://doi.org/10.1016/j.edurev.2014.10.002

Guo, L. (2022a). The effects of self-monitoring on strategy use and academic performance: A meta-analysis. *International Journal of Educational Research, 112*, 101939. https://doi.org/10.1016/j.ijer.2022.101939

Guo, L. (2022b). Using metacognitive prompts to enhance self-regulated learning and learning outcomes: A meta-analysis of experimental studies in computer-based learning environments. *Journal of Computer Assisted Learning, 38*(3), 811–832. https://doi.org/10.1111/jcal.12650

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1–10. https://doi.org/10.1016/j.learninstruc.2016.02.006

Gutierrez de Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology, 114*(4), 681–700. https://doi.org/10.1037/edu0000674

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction, 24*, 37–47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction, 24*, 62–66. https://doi.org/10.1016/j.learninstruc.2012.05.009

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 426–445. https://doi.org/10.1037/1082-989X.9.4.426

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802.https://doi.org/10.2307/2336325

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*(2), 161–176. https://doi.org/10.1007/s11409-009-9042-8

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning, 7*(2), 133–149. https://doi.org/10.1007/s11409-012-9088-x

Kant, J. M., Scheiter, K., & Oschatz, K. (2017). How to sequence video modeling examples and inquiry tasks to foster scientific reasoning. *Learning and Instruction, 52*, 46–58. https://doi.org/10.1016/j.learninstruc.2017.04.005

Kim, J. H. (2018). The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. *Applied Ergonomics, 67*, 193–202. https://doi.org/10.1016/j.apergo.2017.10.006

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 923–954. https://doi.org/10.1037/a0026936

Kok, E., Hormann, O., Rou, J., Van Saase, E., Van Der Schaaf, M., Kester, L., & Van Gog, T. (2022). Re-viewing performance: Showing eye-tracking data as feedback to improve performance monitoring in a complex visual task. *Journal of Computer Assisted Learning, Advance Online Publication.* https://doi.org/10.1111/jcal.12666

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

Kostons, D., van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education, 54*(4), 932–940. https://doi.org/10.1016/j.compedu.2009.09.025

Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Kuhn, J., van den Berg, P., Mamede, S., Zwaan, L., Bindels, P., & van Gog, T. (2022). Improving medical residents' self-assessment of their diagnostic accuracy: Does feedback help? *Advances in Health Sciences Education, 27*(1), 189–200. https://doi.org/10.1007/s10459-021-10080-9

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), 173–194. https://doi.org/10.1007/s11409-010-9056-2

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review*, *35*(4). https://doi.org/10.1007/s10648-023-09819-0

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*(4), 307.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications, Inc.

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

Mihalca, L., Mengelkamp, C., & Schnotz, W. (2017). Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. *Metacognition and Learning, 12*(3), 357–379. https://doi.org/10.1007/s11409-017-9173-2

Mihalca, L., Mengelkamp, C., Schnotz, W., & Paas, F. (2015). Completion problems can reduce the illusions of understanding in a computer-based learning environment on genetics. *Contemporary Educational Psychology, 41*, 157–171. https://doi.org/10.1016/j.cedpsych.2015.01.001

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699–710. https://doi.org/10.1037/a0019182

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

Morrison, J. R., Bol, L., Ross, S. M., & Watson, G. S. (2015). Paraphrasing and prediction with self-explanation as generative strategies for learning science principles in a simulation. *Educational Technology Research and Development, 63*(6), 861–882. https://doi.org/10.1007/s11423-015-9397-2

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology, 33*(6), 1068–1079. https://doi.org/10.1002/acp.3548

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

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*(2), 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. *Journal of Educational Research, 95*(3), 131–142. https://doi.org/10.1080/00220670209596583

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Volume V*)*. OECD Publishing Pisa.

Osterhage, J. L. (2021). Persistent miscalibration for low and high achievers despite practice test feedback in an introductory biology course. *Journal of Microbiology & Biology Education, 22*(2), e00139-e121. https://doi.org/10.1128/jmbe.00139-21

Oudman, S., Van De Pol, J., & Van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning, 17*, 213–239. https://doi.org/10.1007/s11409-021-09281-9

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review, 28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Pilegard, C., & Fiorella, L. (2016). Helping students help themselves: Generative learning strategies improve middle school students' self-regulation in a cognitive tutor. *Computers in Human Behavior, 65*, 121–126. https://doi.org/10.1016/j.chb.2016.08.020

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Elsevier.

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

Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? *Meta-Analytic Insights. Educational Psychology Review, 32*(4), 917–949. https://doi.org/10.1007/s10648-020-09558-6

Raaijmakers, S. F., Baars, M., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2018). Training self-assessment and task-selection skills to foster self-regulated learning: Do trained skills transfer across domains? *Applied Cognitive Psychology, 32*(2), 270–277. https://doi.org/10.1002/acp.3392

Raaijmakers, S. F., Baars, M., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning, 14*(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics, 20*(1), 18–41. https://doi.org/10.4219/jaa-2008-869

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*(6), 1004–1010. https://doi.org/10.3758/bf03209348

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*(4), 262–270. https://doi.org/10.1016/j.learninstruc.2011.10.007

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford University Press.

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology, 46*(2), 168–175. https://doi.org/10.1177/0098628319834381

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. https://doi.org/10.1037/a0021705

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS ONE, 9*(7), e98663. https://doi.org/10.1371/journal.pone.0098663

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*(5), 918–929. https://doi.org/10.1002/acp.3556

Schoenfeld, A. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education, 10*, 173–187. https://doi.org/10.2307/748805

Schoenfeld, A. H. (2015). How we think: A theory of human decision-making, with a focus on teaching. In S. J. Cho (Ed.), *The Proceedings of the 12th International Congress on Mathematical Education* (pp. 229–243). Springer International Publishing. https://doi.org/10.1007/978-3-319-12688-3_16

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making, 18*(1), 29–53. https://doi.org/10.1002/bdm.486

Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison, and validation of meta-essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods, 8*(4), 537–553. https://doi.org/10.1002/jrsm.1260

Sweller, J. (2023). The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review, 35*(4), 95. https://doi.org/10.1007/s10648-023-09817-2

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *Quarterly Journal of Experimental Psychology, 65*(7), 1376–1396. https://doi.org/10.1080/17470218.2012.656665

Testa, I., Colantonio, A., Galano, S., Marzoli, I., Trani, F., & di Uccio, U. S. (2020). Effects of instruction on students' overconfidence in introductory quantum mechanics. *Physical Review Physics Education Research, 16*(1), 010143. https://doi.org/10.1103/PhysRevPhysEducRes.16.010143

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1267–1280. https://doi.org/10.1037/0278-7393.31.6.12

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among seventh and eighth graders. *Journal of Educational Psychology, 104*(3), 554–564. https://doi.org/10.1037/a0028660

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*(2), 264–273. https://doi.org/10.1348/135910710x510494

Van De Pol, J., Van Den Boom-Muilenburg, S. N., & Van Gog, T. (2021). Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning. *Metacognition and Learning, 16*(3), 769–799. https://doi.org/10.1007/s11409-021-09268-6

Van Gog, T., Hoogerheide, V., & Van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review, 32*(4), 1055–1072. https://doi.org/10.1007/s10648-020-09544-y

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. https://doi.org/10.1007/s10648-010-9134-7

Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., Van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

Van Loon, M. H., & Roebers, C. M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. *Early Childhood Research Quarterly, 53*, 301–313. https://doi.org/10.1016/j.ecresq.2020.05.007

Wang, H.-S., Chen, S., & Yen, M.-H. (2021). Effects of metacognitive scaffolding on students' performance and confidence judgments in simulation-based inquiry. *Physical Review Physics Education Research, 17*(2), 020108. https://doi.org/10.1103/physrevphyseducres.17.020108

Wigfield, A., Eccles, J. S., Mac Iver, D., Reuman, D. A., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology, 27*(4), 552–565. https://doi.org/10.1037/0012-1649.27.4.552

Wijnia, L., & Baars, M. (2021). The role of motivational profiles in learning problem-solving and self-assessment skills with video modeling examples. *Instructional Science, 49*(1), 67–107. https://doi.org/10.1007/s11251-020-09531-4

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Routledge.

Wollenschlager, M., Hattie, J., Machts, N., Moller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology, 44–45*, 1–11. https://doi.org/10.1016/j.cedpsych.2015.11.003

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.