META-ANALYSIS

Let Learners Monitor the Learning Content and Their Learning Behavior! A Meta-analysis on the Effectiveness of Tools to Foster Monitoring

Charlotte Dignath[1] · Reyn van Ewijk[2] · Franziska Perels[3] · Sabine Fabriz[4]

Abstract

Self-monitoring is an integral part of self-regulated learning. Tools that foster learners' monitoring, such as learning journals, portfolios, or rubrics, are supposed to promote self-regulation and to improve performance. The aim of this meta-analysis was to examine the effectiveness of tools designed to foster monitoring on learning-related variables (academic achievement, self-regulated learning, and motivation). As these tools vary greatly in their design and the addressed components, this meta-analysis aims to uncover how such tools should be implemented to foster monitoring most effectively. The results of this meta-analysis, integrating 109 effect sizes with 3492 participants from 32 intervention studies, supported a reactivity effect by revealing a moderate effect size on academic achievement ($d = 0.42$), and low effects on self-regulated learning ($d = 0.19$) and motivation ($d = 0.17$). These effects were moderated by characteristics of the tool and their implementation. Effect sizes were highest for tools that (1) focused on the monitoring of both learning content as well as learning behavior, (2) stimulated metacognitive monitoring, and (3) were implemented in shorter studies. On a descriptive level, higher effects were found in favor of monitoring interventions that included teacher feedback on self-monitoring entries and allowed learners to directly revise their work based on this feedback. The findings show that there is substantial variation across tools, which yield theoretical and methodological implications on how to foster monitoring as important parts of the self-regulation cycle.

To become lifelong learners in a forward-thinking high-tech society, learners need to develop the skills of self-regulated learning (SRL). Research over the last decades

has illustrated the effectiveness of SRL for students' learning, achievement, and motivation in various learning settings (e.g., Dent & Koenka, 2016). Self-regulated learners plan, monitor, and evaluate their learning throughout the learning process (Zimmerman, 2000). However, many students lack strategies to engage in self-regulation processes, which can lead to shallow processing (Casadevante et al., 2021). To activate students' self-regulatory activities, teachers can prompt their learners to apply strategies that initiate planning, monitoring, and reflection (Nückles et al., 2021). To this end, monitoring tools, such as learning journals, rubrics, or portfolio, have been regularly used in educational practice from primary school up through higher education (Panadero & Jonsson, 2020; Schmitz & Schmidt, 2011).

However, little evidence has been presented so far that shows the effectiveness of continual use of monitoring tools on students' learning and performance. Moreover, such tools are very heterogeneous and engage learners in different types of regulation processes, and they may also show differential effects for different groups of learners (Panadero et al., 2016). Therefore, the following questions arise: Are these tools indeed promising for the enhancement of SRL and performance? Which characteristics of such tools are effective for which learner? The aim of this meta-analysis was to investigate the effectiveness of monitoring tools and to identify moderating characteristics of this effect.

Monitoring in the Process of Self-regulated Learning

Most researchers agree that SRL includes cognitive, metacognitive, and motivational processes that interact reciprocally (Panadero, 2017, for an overview). Self-regulated learners metacognitively, motivationally, and behaviorally regulate their learning through monitoring and active adaptation of their strategy use to the specific task execution (Zimmerman, 2000; Winne & Hadwin, 1998). This adaptation takes place throughout the learning process when learners plan, monitor, and evaluate their learning, and the results of such an evaluation affect future planning of upcoming learning cycles (Zimmerman, 2000). In the planning phase of the SRL cycle, a learner chooses a goal to strive for and the strategies necessary to reach this goal. In the monitoring phase, the learners observe themselves during task execution to check whether they are still on the right track to reach the goal. Self-monitoring can refer to one's understanding of learning content as well as observation of one's learning behavior. In the evaluation phase, learners need to self-assess their learning progress and their understanding in order to decide whether the goal is accomplished or whether further steps are needed to reach the goal. The learning behaviors in these two phases—monitoring and evaluation—are very similar and only distinguish themselves with reference to when they take place in the learning cycle.

To describe the metacognitive functioning of monitoring and evaluation in more detail, Nelson and Narens (1994) distinguish an object level where cognitions take place during task execution, and a meta-level which contains a representation of the object level, thus, where cognitions about cognition take place. They characterize *monitoring* as the metacognitive thoughts and feelings that evolve from an information flow from the object to the meta-level (for example, a student realizes that they

are not on the right track to solve the problem). Reversely, *regulation* or *control* depicts the following information flow from the meta- back to the object level. This information flow is based on a person's evaluation of the preceding monitoring and initiates a response (for example, the student adapts their learning behavior). Self-monitoring can take place on a micro-level during the task execution phase of the SRL cycle (Zimmerman, 2000), but can also occur on a higher level when recording the whole SRL cycle from a meta-perspective (Schmitz & Perels, 2011).

Winne and Hadwin's COPES model (Winne, 2001; 2004; Winne & Hadwin, 1998) describes learning processes in even more detail. According to the COPES model, learning occurs in four basic stages, wherein a person's *conditions*, *operations*, *products*, *evaluations*, and *standards* interact. These facets, excluding *operations*, are types of information that learners use or generate during the learning process. *Conditions* are the resources available and the constraints imposed by a task or environment. They affect both a person's standards and operations, and can be internal and external. Whereas cognitive conditions are internal resources of the learner, such as beliefs, motivational orientations, and knowledge, which result from memories of past learning experiences, task conditions are external cues that include instructional cues, time, and context. *Standards* are multi-layered criteria that learners believe represent the optimal end state, and they include both measures (for example, what needs to be learned to complete a task successfully) and beliefs (for example, about the difficulty of the task). Learner use these standards to determine the success of their operations. *Operations* include cognitive information processing, such as searching, monitoring, assembling, rehearsing, and translating. These processes result in cognitive products, such as the recall of information.

By means of *monitoring*, these products are compared with the person's standards to determine if the goals have been met. Monitoring is a cognitive process of comparing characteristics of a particular target to a list of standards that include the attributes of an ideal target (Winne, 2004). *Cognitive monitoring* includes cognitive evaluations, such as calibrations or judgments of learning, which refer to the object level focus of monitoring in the model by Nelson and Narens (1994). When cognitive evaluations show a mismatch between products and standards, adaptations should be made. Such adaptations entail metacognitive control over the choice of learning strategies and processes. *Metacognitive monitoring* is a special case of monitoring and occurs when meta-level information, for example on the actual task difficulty, does not match the previously established standard of task difficulty. Thus, metacognitive monitoring includes the evaluation of the SRL processes. This may activate a metacognitive control strategy whereby this particular standard is amended, and could in turn influence other standards and, eventually, may lead to updates of products from previous phases. *Metacognitive control* means deciding how to act based on an evaluation created through metacognitive monitoring (Winne, 2004).

According to the COPES model, there is no typical learning cycle, but in most learning processes, the cognitive architecture is traversed until there is a clear definition of the task (phase 1), followed by the development of learning objectives and the best plan to achieve them (phase 2). This leads to the use of strategies to start learning (phase 3). Learning products are compared to standards, such as the overall

accuracy of the product or, the learner's ideas about what needs to be learned. If the product does not meet the standard, further learning operations are started. Finally, students may choose to make more significant and longer-term changes in their beliefs, motivational orientations, and strategies that constitute SRL (phase 4).

Mode of Action of Monitoring Tools

Cognitive and metacognitive monitoring play a key role in every phase of the described learning processes of the COPES model (Winne & Hadwin, 1998). Thus, a lack of monitoring of standards from phases 1 or 2 can be just as troublesome as a lack of monitoring of phase 3 products. Sustained monitoring appears to be an integral component of successful SRL (Greene & Azevedo, 2007). Metacognitive and cognitive monitoring are discussed together, given that they both pertain to students' monitoring processes in the learning cycle. However, some important distinctions exist between the two monitoring processes. While cognitive monitoring refers to the subject matter that the learner is dealing with, metacognitive monitoring refers to topics about the subject matter, such as the properties of the cognitive operations learners use (Winne, 2004). Intervention studies to promote cognitive or metacognitive monitoring are often based on different research traditions, depending on whether the focus is on promoting metacognitive or cognitive monitoring.

Tools to Support Cognitive Monitoring

The question arises whether monitoring can be triggered effectively by tools that are designed to result in a reactivity or reminder effect. Existing research on the effectiveness of cognitive and metacognitive monitoring has delivered inconsistent results.

One key element in SRL is that learners need to generate information about their learning and understanding through cognitive monitoring. To this end, learners compare their performance to an internal or external standard (Winne, 2004). Based on the outcome of this comparison, the learners will evaluate their performance, and this evaluation has implications on the learners' further learning activities (Butler & Winne, 1995). For example, a learner who evaluates performance as correct and the learning goal as reached may stop working on this assignment. This self-assessment draws on the results of the learner's self-monitoring as well as on the quality of the learner's identified standard (Winne, 2004). If the identified standard is of poor quality, self-assessment can be biased or inaccurate (Brown et al., 2015).

Research on *cognitive monitoring*, such as calibrations or judgments of learning, showed that these cognitive evaluations can affect both performance and metacognitive monitoring (Stone, 2000). Some studies indicated that self-assessment helps improve self-regulation skills (Ramdass & Zimmerman, 2008), strategy use (Brookhart et al., 2004), and motivation (Olina & Sullivan, 2002), but the evidence for a relationship between self-assessment and SRL remains inconsistent (Brown & Harris, 2013). In order for students to properly assess their learning, they must have

both an appropriate learning product from phase 3 and a correct task definition or standard to which they can compare this learning product (Winne, 2004). Beside the determination of such assessment criteria that form the standard, cognitive monitoring requires learners to seek internal or external feedback. External feedback can be found by asking teachers or peers directly, or by observing the learning evidence. Internal feedback can be obtained by observing one's internal states, such as emotions. Drawing on this external and internal feedback, learners reflect on their learning and derive a self-assessment judgment (Yan & Brown, 2017).

To investigate and to enhance learners' prediction accuracy and metacomprehension judgment, learners are asked to self-assess their learning and performance. Yet, research indicated that many students are poor at correctly assessing their learning and performance, and that those students who performed the worst on a test were the most likely to misjudge their performance and to overestimate themselves (Brown & Harris, 2013; Greene & Azevedo, 2007). Self-assessment judgments have been found more accurate when the comparison standards were specific, when students were trained explicitly in self-assessment, and when they were provided with feedback (Brown & Harris, 2013; Panadero et al., 2016).

Research has produced a large variety of implementations of self-assessment tools in the classroom that differ in complexity (Panadero et al., 2017), such as rubric-guided judgments, self-rating, or self-estimates of performance (Brown & Harris, 2013). For example, rubrics aim to foster self-assessment by providing learners with external standards and criteria to support their evaluative judgment (Andrade et al., 2008). Another way to promote self-assessment are self-ratings that focus the learner's attention to evaluate the quality or quantity of their work using a grading system. (Baleghizadeh & Masoun, 2013). Possible self-rating prompts are "I am pleased with my work because I…" or "I would grade myself A B C D E because I…" (Clarke et al., 2003). More simple self-rating practices like self-marking can be implemented by means of a marking guide for objectively answered questions (Todd, 2002). Another alternative to self-assessment is for students to assess their level of performance or ability against a test or task. Possible prompts can be "How well have I done on this test?" (Brown & Harris, 2013). Finally, a learning portfolio can be used as a self-assessment tool when its main purpose is to assess one's learning outcomes (Chang et al., 2013). Portfolios are implemented to collect artifacts of student learning that should help students to reflect on their progress and to foster self-assessment (Abrami et al., 2013). Portfolio assessment is characterized by a systematic collection of student work that documents a student's efforts, progress, and achievements (Chang, 2008). They usually contain a student's profile, the setting of learning goals, uploading of assignments and artifacts, and reflective writing or self-scoring (Chang et al., 2013). The prompts in portfolios are similar to those in other self-assessment tools, but, additionally, the portfolio is designed to support the evaluation process by collecting and uploading artifacts.

Tools to Support Metacognitive Monitoring

In addition to cognitive monitoring, leaners can monitor on a metacognitive level (Winne, 2004). Research showed that high ability of *metacognitive monitoring* is

associated with better performance (e.g., Lan, 1996), but that most students cannot monitor their learning effectively (Stone, 2000). Little attention has been paid to training learners in the ability of metacognitive monitoring, but the few studies undertaken showed promising results (e.g., Delclos & Harrington, 1991; Lan, 1996). Despite the high effectiveness, when learners feel overwhelmed by the task, they were found to prefer cognitive monitoring over metacognitive monitoring (Winne, 2001).

Tools to foster metacognitive monitoring aim to activate self-observation during the learning process by focusing the learner's attention on their understanding or on their learning behavior (Zimmerman, 1989). More specifically, such tools provide the learners with awareness and self-generated feedback about their own comprehension and their own performance (Butler & Winne, 1995). Thereby, learners can identify discrepancies between the current state and the desired state of learning, which enables them to realize whether additional effort is needed. Most tools include questions or prompts intended to stimulate metacognitive monitoring by pointing learners' attention to their learning process or understanding (Ferreira et al., 2015). This shift of attention focus alone is already assumed to elicit behavior change, also called the *reactivity effect* (Zimmerman, 2002). Specific self-monitoring techniques have been found to enlarge this effect. For example, a learning journal that includes self-recording requires learners to register their actions during task performance (e.g., Schmitz & Perels, 2011, p. 271: "I managed to realize my intentions today"; Zimmerman & Moylan, 2009). In the same vein, providing students with prompts for monitoring one's understanding is supposed to encourage self-monitoring during the learning process (e.g., Nückles et al., 2010, p. 241: "Which main points haven't I understood yet?").

A second function of metacognitive monitoring regards the *reminder effect* (Webber et al., 1993). The questions in a metacognitive monitoring tool can work as cues that illuminate the relevance of a certain topic in the moment that the learner works with the tool (Schmitz & Perels, 2011). For example, being asked to describe one's plans for the homework session drives the learner to reflect about his or her plans. In this way, the tool does not only focus the attention to a certain behavior but also prompts the use of a certain strategy (Nückles et al., 2021). Thus, beyond fostering metacognitive monitoring, some tools contain prompts that stimulate strategy use and, therefore, engage students not only in monitoring but also in control processes.

Evidence on the Effectiveness of Monitoring

Evidence from the Field of Formative Assessment

Self-assessment lies close to formative assessment as it calls for learners' active participation in the evaluation of their own work and its comparison against a certain standard. As in formative assessment, self-assessment implies a "growth mindset" by enabling learners to improve their work, emphasizing that learning is incremental and not simply understanding vs. not understanding (Sanchez et al., 2017). When self-assessment is used in a formative way, it can serve as a learning strategy, helping students to evaluate their progress (Yan & Brown, 2017).

The literature review by Black and Wiliam (1998) summarizes research on formative assessment, showing highly positive effects of formative assessment on students' academic achievement. Yet, Black and Wiliam (1998) concluded that students are not necessarily able to benefit from formative assessment unless they metacognitively comprehend their own understanding in order to undergo conceptual change. Kingston and Nash (2011) found only a lower average effect size ($d = .28$) in their meta-analysis on the effect of formative assessment on academic achievement. Moderator analyses indicated that effect sizes differed mainly due to the school subject, with English and language arts producing the highest effect sizes and mathematics and science leading to the lowest effects for formative assessment.

For the field of writing, the meta-analysis by Graham et al. (2015) produced an average effect size of formative assessment on students' performance ($d = .63$), with the highest effect sizes when feedback was provided by teachers ($d = .89$). Medium effect sizes were found for studies in which students were taught to self-assess their writing ($d = .62$) and for studies in which students received feedback from their peers ($d = .62$). The lowest effect sizes resulted from studies that tested the effects of computer feedback on writing performance ($d = .38$). In line with Graham's results, the analysis of 195 studies in Hattie's meta-synthesis (2016) revealed a moderate effect of formative evaluation on student learning ($d = .68$).

Similar to self-monitoring and self-assessment, formative assessment aims to provide learners with feedback about their progress. However, in formative assessment the teacher monitors the students' learning and provides them with external feedback, whereas in self-assessment the learners have to self-monitor and obtain external as well as internal feedback about their progress.

Evidence from the Field of Self-assessment

Two meta-analyses examined associations of self-assessment and test performance. Sitzman et al. (2010) computed meta-analytic correlations of adult learners' self-assessment with achievement, motivation, and self-efficacy, which produced large effects. Their results revealed a large effects regarding achievement ($r = .34$; corresponds to $d = 0.72$), and even larger with self-efficacy ($r = .43$; corresponds to $d = 0.95$), and motivation ($r = .59$; corresponds to $d = 1.46$). Sanchez et al. (2017) computed meta-analytic correlations of primary and secondary school students' self-grading with grades assigned by teachers and found an even higher correspondence ($r = .67$; corresponds to $d = 1.81$).

In addition, several meta-analyses have been conducted in recent years on the effectiveness of self-assessment interventions in school- and higher education settings, which overall yielded medium effect sizes. A meta-analysis by Panadero et al. (2017) included studies from primary school through higher education and displayed small to medium effects of self-assessment on SRL ($d = 0.23$), but large effects on self-efficacy ($d = 0.73$). With regard to younger learners, Brown et al. (2012) reviewed research on the effects of self-assessment practices in kindergarten through grade 12. They did not report a mean effect size for their systematic review, but identified effect sizes on achievement ranging from $d = -0.04$

up to $d = 1.62$, with a median effect size around $d = 0.40$. Sanchez et al. (2017) investigated the effectiveness of self-grading on subsequent test performance in primary and secondary school and found a medium effect size ($g = 0.34$). In a recent meta-analysis, Yan et al. (2021) found self-assessment interventions in higher education to successfully improve achievement ($g = 0.45$). Effects were significantly higher in interventions involving explicit feedback ($g = 0.66$) than in those without feedback ($g = 0.21$).

Evidence from the Field of Self-monitoring

Comparable to the inconsistent evidence for formative assessment and self-assessment, studies on the effect of self-monitoring tools have produced inconsistent results. For learning journals with open answering format, the meta-analysis by Bangert-Drowns et al. (2004) on the effectiveness of writing-to-learn interventions revealed only a small effect of journaling on academic achievement ($d = 0.20$) overall, but they found that the use of metacognitive prompts significantly increased this effect ($d = 0.26$). A more recent meta-analysis by Guzman et al. (2018) found higher effects of self-monitoring on reading performance (Tau-$U = 0.79$; corresponds to $d = 0.47$). Looking more generally at the effects of interventions to foster monitoring of goal progress, a meta-analysis by Harkin et al. (2016) revealed positive effects of progress monitoring on goal attainment ($d = .40$). In a recent meta-analysis, Gutierrez de Blume 2022 investigated the effectiveness of learning strategy interventions on metacognitive monitoring accuracy. His findings indicate that learning strategy instruction has a medium effect on monitoring accuracy ($g = 0.56$), which was significantly higher with adult-only samples. Although in this meta-analysis, the intervention was not directly targeted at supporting monitoring but by providing learning strategy instruction, the findings demonstrate that metacognitive monitoring can be supported by educational intervention.

Taking this evidence into account, monitoring tools may have promising effects for learning with mean effect sizes around $d = 0.40$. However, the inconsistent results suggest that certain characteristics may influence the effectiveness of these tools and call for further systematic research.

Potential Variables that May Moderate the Effectiveness of Monitoring Tools

Based on the mode of action described earlier, we derive five characteristics of the implementation of monitoring tools that may affect their effectiveness: (a) the type of monitoring stimulated by the tool; (b) the focus of the tool on learning content and learning behavior; (c) whether learners receive teacher feedback on their entries; (d) the duration of the intervention; and (e) the age of the participants.

The Focus of the Tool

Monitoring tools vary in terms of their focus. Some tools focus on learning content, whereas others focus on learning behavior, and some tools focus on both. For example, some learning journals have to be completed by answering questions in an open-answer format meant to prompt metacognitive reflection in learners through reflective writing about the learning process (e.g., Nückles et al., 2009), whereas others are semi-structured or highly structured, like a questionnaire that consists of scales or items to assess the learner's use of SRL strategies (e.g., Costa Ferreira et al., 2015). By answering these questions, the attention should be drawn to the learner's current situation in order to encourage self-monitoring, self-assessment, or self-reflection and, therefore, to stimulate self-regulation (Glogger et al., 2012). Compared to such highly structured tools, learning journals with an open answering format are often more closely linked to the learning content (Bangert-Drowns et al., 2004). These tools are sometimes also called portfolio and serve the learners to collect artifacts of their learning process. These artifacts provide evidence for learning and can consist of assignments, essays, projects, presentations, and other media. In this way, the portfolio makes the learning progress visible and can help the learner to reflect on this progress (Paulson et al., 1991). Several studies revealed, however, that reflective writing without any additional guidance is less beneficial for SRL than if prompts are given to guide the learner towards what is important (Nückles et al., 2021).

The Type of Monitoring Stimulated by the Tool

Tools to support monitoring vary substantially in the way they activate learning. This is shaped by the way how monitoring is stimulated. In the retrieved primary studies, we distinguished three categories that were inductively derived from the sample of coded primary studies: studies which were embedded in stimulated cognitive evaluations which mainly concern the object level (e.g., Andrade & Valtcheva, 2009), metacognitive evaluations predominantly concerning meta-level information (e.g., Zimmerman, 2000), and studies that activated monitoring indirectly by requiring students to collect artefacts of their work (portfolio tools, e.g., Paulson et al., 1991). The idea of most tools that stimulate cognitive monitoring is to specifically address the standards that learners should judge their performance against, for example, by providing a list of criteria (Panadero et al., 2016). By this, these tools focus the learner systematically to the comparison of their own understanding and performance against a goal standard by providing a structure of this standard. The collection of artifacts are used to encourage learners to reflect on their learning evidence in order to enhance cognitive self-evaluation. Thus, they focus the learners' attention to the learning products rather than on the standards to compare these products with. Although there was some overlap as several studies draw on cognitive and metacognitive aspects of monitoring (e.g., Panadero et al., 2013) or encourage cognitive evaluations and the collection of artefacts at the same time (e.g., Güzeller,

2012), for every study it was possible to assign the monitoring tool to one of the three categories.

External Feedback

Many researchers argue that to develop effective monitoring, learners can be supported by external feedback (Yan & Carless, 2021). Feedback provides learners with evaluative information about their progress, which scaffolds monitoring and thereby enhances SRL. When students generate internal feedback, they compare their performance or understanding with their standard and may identify a discrepancy between the goals and the product. This feedback ideally informs learners whether any action is required to reduce this discrepancy (e.g., Butler & Winne, 1995). However, many learners are poor at self-assessment and therefore may not identify the discrepancy between goal and outcome, or they may set inappropriate goals, and, as a result, will not regulate appropriately (Chou & Zou, 2020). Therefore, researchers have proposed to provide learners with external feedback to support them in evaluating their learning progress and in deciding about methods to regulate learning (e.g., Hattie & Timperley, 2007). Moreover, external feedback can help learners establish a realistic internal standard to compare their performance to (Brown et al., 2012; Panadero et al., 2020).

Feedback can take place on several levels (Hattie & Timperley, 2007). Feedback on the learning outcome states whether the tasks are solved correctly or not. This helps learners to detect potential discrepancies between the learning product and the standard, but does not inform on how to resolve this discrepancy. By contrast, the goal of process feedback is to guide learners to find and correct their mistakes. Besides monitoring information, process feedback informs about adequate regulatory actions to reduce the discrepancy. On a meta-level, self-regulation feedback informs the learners whether their self-assessment is correct and whether or not it needs to be adjusted. Thus, self-regulation feedback no longer refers only to the learning content and progress but, on a meta-level, to the monitoring and regulation of learning that has taken place. Depending on the level of feedback, external feedback can scaffold learners' development in cognitive and metacognitive monitoring and regulation. Several meta-analyses in this field have investigated whether external feedback moderates the effects of self-assessment and self-monitoring on academic achievement. With regard to self-assessment, Sitzmann et al. (2010) found the learning outcomes of adult learners to be stronger in courses that provide external feedback on learners' performance ($p = .28$) than in courses without feedback ($p = .14$). Yan et al. (2021) also found significantly higher effect sizes in interventions in higher education, in which learners received external feedback, than in interventions without feedback. This is in line with findings from the meta-analysis by Graham et al. (2015), showing more specifically that teacher feedback ($d = .89$) exceeded self- and peer feedback (both $d = .62$), as well as computer feedback ($d = .38$) in regard to learning outcomes of primary and secondary school students. However, in the meta-analysis by Bangert-Drowns (2004), feedback was not found to moderate the effects of self-monitoring interventions. For example, Raaijmakers et al. (2019)

did not find any beneficial effect of feedback on subsequent self-assessment accuracy, concluding that their self-assessment feedback might have led learners to pay less instead of more attention to their self-assessment. Wong et al. (2019) suggested that learners' individual differences affect whether they can benefit from external feedback. Moreover, it could also be that the different levels of feedback have different effects on learning (Chou & Zou, 2020). Differential effects depending on both learner characteristics and on feedback level could account for the inconsistent findings in earlier research.

Frequency and Duration of the Intervention

How often a monitoring tool is used is likely to influence the intensity of the intervention. Their repeated application is part of the assumed strength of these tools and keeps learners vigilant in their learning process. For example, a study by Ziegler (2014) on the effectiveness of the European Learning Portfolio suggests a better effect with increased frequency of use. Learning journals are usually applied once or twice a week (Tezci & Dikici, 2006; Wäschle et al., 2015), or daily when used as an additional measurement to capture SRL in the process (Schmitz & Perels, 2011; Stoeger et al., 2015). Although one would assume a learning effect with increasing practice, the meta-analysis by Sitzmann et al. (2010) and by Bangert-Drowns (2004) did not reveal effects of self-assessment or journaling to differ as a function of the frequencies of intervention. A linear relation is unlikely because intervention studies with learning diaries often report motivational decline over time (Dignath et al., 2015; Fabriz et al., 2007).

Age of the Participants

While earlier research on SRL had considered young children up to secondary school age to be unable to coordinate the metacognitive processes of SRL (e.g., Myers & Paris, 1978), there is empirical evidence from the last two decades showing that already very young children are able to self-regulate to a certain extent (Whitebread et al., 2007), and that young children can benefit from SRL support (Perry & Rahim, 2011). More precisely, monitoring skills have found to develop between the age 7 to 10 on (Roebers et al., 2011). Some authors have highlighted the appropriateness of using tools to foster monitoring, such as learning journals, for different age groups (Glaser & Brunstein, 2007; Klug et al., 2011). Whereas some meta-analyses did not find effect sizes to vary by educational level (e.g., Kingston & Nash, 2011 on formative assessment; Panadero et al., 2017 on self-assessment), other meta-analyses found effects of self-assessment to increase with age (e.g., Sanchez et al., 2017). At the same time, the meta-analyses on the effectiveness of SRL training by Dignath and Büttner (2008) and by Hattie et al. (1996) found that younger students benefitted from different training characteristics just as much as older ones. Given the unclear findings, the impact of the learners' age should be taken into account.

Aim of this Study

The use of monitoring tools, such as learning journals, portfolio, or rubrics, has been widely established within schools and in higher education. Ministries of education and institutions of teacher education advise teachers to apply such tools in order to improve learning behavior, motivation, and achievement (Clark, 2012). Yet, little empirical evidence exists on which type of tool is most beneficial for whom. With this meta-analysis, we hope to contribute to the understanding of tools to promote monitoring in school and higher education. The scope of this meta-analysis includes studies in which educational tools were used to stimulate learners' cognitive or metacognitive monitoring.

Intervention research on promoting monitoring has been conducted within separate research traditions, although tools that target learners' cognitive or metacognitive monitoring are closely linked (Panadero et al., 2018). Our meta-analysis extends previous research by updating the evidence base and by combining similar interventions from different research traditions. The main goals of this meta-analysis are to investigate whether and when monitoring tools are effective techniques to foster learning. Meta-analysis can be used to combine studies investigating the same question and determine a mean effect. In addition, it can also be used to test the extent to which different characteristics of the studies influence this effect. In this meta-analysis, we integrated studies that examine the effectiveness of monitoring tools. In doing so, we also test whether efficacy differs based on the type of monitoring (cognitive vs. metacognitive monitoring vs. indirect stimulation via collecting artifacts), by including the type of monitoring as a moderator variable in our analysis.

To enlarge our understanding of the mechanism of how such tools affect SRL, motivation, and achievement, the first aim of this study is to test the effect of interventions with self-monitoring and self-assessment tools on (1) academic achievement; (2) the use of SRL strategies; and (3) learning motivation. Second, looking at the present evidence, it is unclear which characteristics of monitoring tools and their implementation are (most) effective and whether this differs for different target groups. Thus, our second aim is to answer the question of how these tools can be implemented most effectively. Therefore, the following research questions were investigated:

1. Overall Effectiveness of Monitoring Tools

Do monitoring tools have an effect on learners' academic achievement, SRL, and motivation? In terms of a practically significant effect, we expect the effect size to be at least $d = 0.40$ because this was the average effect found for any kind of educational intervention, which can serve as a benchmark (Hattie, 2012).

H1: We expect an effect size of $d = 0.40$ or more for achievement, SRL, and motivation.

2.　Variables That Might Moderate This Relationship

2.1 Does the effect differ depending on the focus of the tool on learning content and/or learning behavior? Interventions to foster SRL were found to be most effective when being connected to the learning content (Hattie et al., 1996). This facilitates the learners' transfer from the intervention to the real classroom (Wang & Sperling, 2020). As monitoring activities comprise both monitoring one's understanding of the learning content as well as monitoring one's learning behavior (Pressley & Gathala, 1990), it can be assumed that monitoring tools are more effective when they do not focus solely on either the monitoring or evaluation of one's learning content or one's learning behavior, but on both simultaneously.

H2.1: We expect the largest effects for tools that focus on both learning content and learning behavior at the same time.

2.2 Does the effect vary as a function of the type of monitoring stimulated by the tool? As tools for metacognitive monitoring additionally integrate planning, monitoring, and evaluation activities of the learning process, we assume that tools stimulating metacognitive monitoring yield larger effects than tools just stimulating cognitive monitoring or tools stimulating the collection of artifacts and therefore activate monitoring only indirectly.

H2.2: The effect is larger for interventions using tools to stimulate metacognitive monitoring than using tools to stimulate cognitive monitoring or collection of artifacts.

2.3 Can the effect be boosted by teacher feedback? External feedback can support the learners' monitoring and evaluation by providing them with evaluative information about their learning (Butler & Winne, 1995; Panadero et al., 2020). Three meta-analyses found external feedback to positively moderate the effects of self-assessment (Sitzman et al., 2010; Yan et al., 2021) and formative assessment (Graham et al., 2015). Even though some studies did not find positive effects of feedback (e.g., Raaijmakers et al., 2019), the large evidence base in favor of feedback suggests that external feedback can boost the effect of monitoring tools.

H2.3: We expect larger effects when learners receive teacher feedback on their entries in the tool.

2.4 Does this effect increase with the duration of the intervention? The meta-analysis by Dignath and Büttner (2008) revealed that longer interventions to promote SRL were more effective. Learners need time to understand and implement SRL strategies, and longer intervention provides more opportunities to practice strategy use. Thus, one can assume that monitoring tools have larger effects when learners have more opportunities for deliberate practice throughout a longer intervention.

H2.4: We expect larger effects with an increasing duration of intervention.

2.5 How is this effect moderated by the age of the learners? In a meta-analysis on SRL training for primary school students, Dignath et al. (2008) found that SRL interventions were more effective to foster the younger primary students' SRL and their motivation. Yet, students' age did not moderate the training effect on academic achievement (Dignath et al., 2008). When comparing the effects of SRL training for primary and secondary school learners, Dignath and Büttner (2008) found higher effects on SRL for older learners, but higher effects on motivation for younger learners. For effects on academic achievement, no clear pattern was found (Dignath & Büttner, 2008). In their meta-analysis on the effectiveness of strategy training, Donker et al. (2014) did not find learners' age to moderate the effects. Likewise, in a recent meta-analysis on SRL training, Wang and Sperling (2020) did not find learners' age to be a moderator of the training effect. Finally, in the meta-analysis by Sanchez et al. (2017), learner age also did not moderate the effect of self-grading. These findings from previous research do not suggest any age differences.

H2.5: We do not expect a moderator effect for students' age.

In addition to the quantitative analyses of the meta-analysis, we also explored these research questions in more depth through qualitative exploratory analyses.

Methods

Eligibility Criteria

In order to identify studies that provide evidence on the effectiveness of monitoring tools, studies that investigated the effects of using such tools on learning outcomes, SRL, or motivation with a quasi-experimental design were sought. Studies were included in the meta-analysis after meeting the following eligibility criteria:

1. Publication type: As the exclusion of grey literature from meta-analyses can lead to exaggerated estimates of intervention effectiveness (McAuley et al., 2000), we included original research that was published in peer-reviewed journals, but also unpublished studies, such as dissertations or reports, and did not restrict publications to ranked journals.
2. Language: Studies published in English were included.
3. Year of publication: Our literature search was conducted at the beginning of 2021, thus allowing studies published by the year 2020 to be included.
4. Population type: The study should be conducted in an educational setting (primary, secondary, tertiary education, vocational education, special education) and therefore involved students from the age of school enrollment through adulthood.
5. Intervention type: The treatment should aim at fostering monitoring in a learning context by using a monitoring tool, such as a learning journal, rubrics, or portfolio.
6. Duration of intervention: The study should investigate regular usage of the tool, i.e., one or more times per week for a minimum of three times. Therefore, experimental studies examining the effects of metacognitive prompts or self-assessment

tools in a laboratory setting in one single session were excluded from the meta-analysis (e.g., Berthold et al., 2007, had to be excluded due to a one-time measurement).

7. Study design: Strict methodological criteria were applied in order to only include studies with high methodological standards. Concerning the study design, the sample should consist of at least 10 students per group so that parametric statistical procedures were applicable. Therefore, no case studies or studies with less than 10 participants per group were included (e.g., Matas & Allan, 2004, could not be included due to the small-sample size).

Moreover, only those studies that had both a pre-post design and a control group design were included in the meta-analysis. The reason for the choice of a control group design was that studies using a pre-post single-group design investigate change within a person relative to the variability of change scores instead of the variability within groups. Studies with less sophisticated designs investigate different research questions and were therefore excluded from this meta-analysis (e.g., Güven et al., 2014).

Beside the control group design, we required the studies to have a pre-post design. This is as all studies followed quasi-experimental designs and were conducted within natural (university) classroom settings, so no randomization was possible. If a randomization took place, this concerned only the assignment of whole classrooms to the conditions, but not of participants to the conditions. Therefore, we excluded studies with a single-group design (e.g., Tican & Taspinar, 2015).

Finally, studies that tested two interventions against each other without any control group comparison (i.e., a second condition without a monitoring tool but with an alternative intervention) could also not be included for reasons of comparability (e.g., Nückles et al., 2010, could not be included as they compared two learning journal conditions), although the methodological quality of such studies with alternative treatment is higher than for studies without any control condition.

8. Types of outcome measures: Studies had to test effects on academic achievement, SRL, or motivation as outcome variables. Studies that reported outcomes of SRL referred to learners' use of strategies to regulate their cognition, metacognition, or motivation during learning. Motivational outcomes included motivational constructs, such as the learners' self-efficacy or their learning motivation, but not the regulation of motivation strategies (this would be classified as SRL outcome). In studies applying comprehensive questionnaires to assess SRL and motivation with one single instrument (e.g., the MSLQ; Pintrich et al., 1991), we extracted the subscale assessing SRL to compute the effect size for SRL, and included the subscale measuring motivation in the effect size representing motivation.

Retrieved studies that did not meet all these criteria were excluded from the meta-analysis.

Literature Search

A literature search was carried out in the common databases used in the field of educational psychology (PsycInfo, PsycArticles, ERIC, and Web of Science) to identify studies that fit into the scope of this meta-analysis. The following keywords were used to define the search field: *self-monitoring; self-assessment; learning diary/diaries*; *learning journal*; *journal writing*; *logbook*; *learning protocols*; *reflective diary/diaries*; *reflective journal/journaling*; *portfolio*. The search produced 1038 results in PsycInfo, ten in PsycArticles, 2,127 in ERIC, and 1777 in Web of Science. After duplicates had been removed, 3987 studies were screened for eligibility based on title and abstract, revealing 120 full-text articles. In order to avoid the inclusion of duplicate data, articles that reported the same data as another article were excluded. An in-depth coding of these articles excluded 33 articles that either reported qualitative studies (e.g., Fazio, 2001), were overview articles on the topic, or provided no data. Thirty-eight studies were excluded due to lack of a control group (e.g., Beckers et al., 2019), pretest (e.g., Andrade et al., 2008), or posttest data (e.g., Cohen Goodman, 1998), or that provided no control group without treatment (e.g., Khodadady & Khodabakhshzade, 2012). Moreover, seven articles had to be excluded because the treatment time was too short to meet our eligibility criteria (e.g., Panadero & Romero, 2014). If articles did not report mean values, standard deviations, or sample size, authors were contacted in order to get the missing data. Several studies had to be excluded due to the required data not being provided. Further studies were removed because of data sources that were repeatedly reported, not meeting the educational context criteria, or similar reasons. As a final result of the study selection procedure, 29 studies met the eligibility criteria and could be included in the meta-analysis. From these studies, 109 effect sizes were extracted (see Fig. 1).

Coding Scheme

We developed a coding scheme, following instructions for systematic coding in order to extract data accurately from each study (Lipsey & Wilson, 2001). A coding training was conducted that included an introduction into the coding scheme, communal coding, and discussion of the coding results among the authors. The first and second author conducted the initial screening of the abstracts together. After the selection procedure, the first author coded all full-text articles, with one of the other two authors randomly coding studies in parallel in order to check intercoder reliability. Intercoder agreement was satisfying (Cohen's kappa = .96; Cohen, 1960). Disagreement was resolved by discussion. If coding information was not available from the article, the authors were contacted in order to retrieve the missing information. The following table provides the categories used for general information, study-specific characteristics, and statistical information about group differences that were coded for each study. It also displays aggregated information on the included studies (see Table 1).
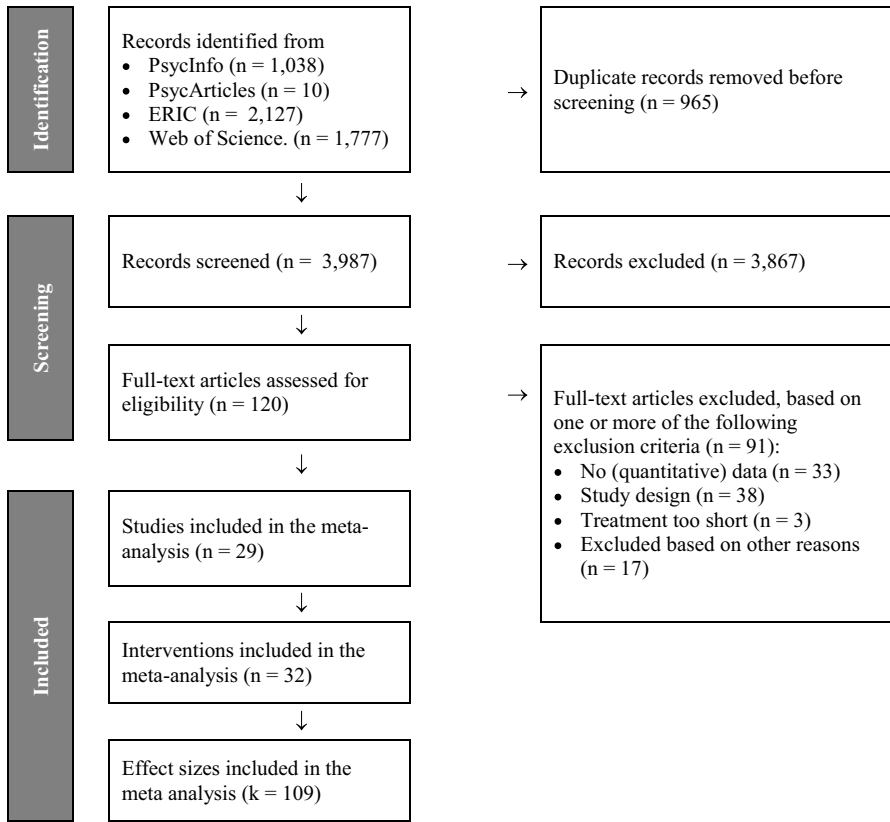
Fig. 1 PRISMA flow chart of the systematic search

Calculation, Weighting, and Adjustment of Effect Sizes

We calculated effect sizes for each study in order to allow for comparing several measures in a standardized way. Since the included studies followed a pre-post control group design, mean differences between pretest and posttest, as well as between intervention and control group, were identified. Effect sizes were calculated following (Morris, 2008) as follows:

$$d = \left(1 - \frac{3}{4(N_T + N_C - 2) - 1}\right) \frac{\left(\overline{Y}_{T,\text{post}} - \overline{Y}_{T,\text{pre}}\right) - \left(\overline{Y}_{C,\text{post}} - \overline{Y}_{C,\text{pre}}\right)}{\text{SDpooled}_{\text{pre}}}$$

in which N_T and N_C are the sample sizes for the treatment and control group, respectively; $\overline{Y}_{T,\text{post}}$ and $\overline{Y}_{T,\text{pre}}$ are the post- and pretest sample means for the treatment group, and $\overline{Y}_{C,\text{post}}$ and $\overline{Y}_{C,\text{pre}}$ are the same for the control group. The pooled standard deviation was calculated as follows:

Table 1 Coded information for each study with aggregated descriptives for included studies

General Information for Each Study	
Publication year	Year of publication as a continuous variable, ranging from 1992 to 2022
	Considering periods of 5 years, most studies were conducted between 2011 and 2015 ($n = 11$)
Country	Country in which the study was conducted (Germany: $n = 6$, Turkey: $n = 6$, USA: $n = 5$, Canada: $n = 2$, Iran: $n = 2$, Spain: $n = 2$, Belgium/Hongkong/Jordan /Kazakhstan/Portugal/Romania: each $n = 1$)
Journal	Journal in which the study was published
Publication type	Journal article/book chapter/conference paper/ thesis/research report (Journal articles: $n = 27$, Dissertation theses: $n = 2$)
Study-specific characteristics	
Design of the study	Pre-post control group design (no/yes) as an eligibility criterion
Aim of the study*	Evaluation of the effect of using a monitoring tool on one of the three outcome categories (academic achievement: $k = 24$, SRL: $k = 53$, motivation: $k = 32$)
Focus	Focus of the tool (learning content: $n = 13$, learning behavior: $n = 8$, learning behavior and learning content: $n = 8$)
Theoretical background	Main focus of the theory section (SRL: $n = 17$, self-assessment: $n = 6$, portfolio: $n = 6$)
External feedback	Feedback from teacher about learners' entries in the tools (no: $n = 18$, yes: $n = 11$)
Duration	Length of intervention period with monitoring tool (in weeks) ranging from 2 to 28 weeks ($M = 12.43$, $SD = 6.92$)
Age	Approximate age of participants in years, ranging from 9 to 26 years ($M = 17.22$, $SD = 4.91$)
Format	Response format of the monitoring tool (closed: $n = 16$, open: $n = 16$, mixed: $n = 3$)
Statistical information about group differences	
Means and standard deviations intervention group (using a monitoring tool) pre and post	
Means and standard deviations control group pre and post	

*Number of effect sizes, as within the studies, more than one outcome can be in the focus

$$\text{SDpre}_{\text{pooled}} = \sqrt{\frac{\left((N_t - 1) * \text{SDpre}_T + (N_c - 1) * \text{SDpre}_C\right)}{N_t + N_c - 2}}.$$

In all analyses, estimates were weighted by the inverse of their sampling variance (Morris, 2008).

Computing a Weighted Mean Effect Size

In order to calculate the average effect of monitoring tools, an overall mean effect size was computed that represents the average effect on learning. Moreover, mean effect sizes were combined for each of the three different outcome groups (academic achievement, use of SRL strategies, motivation) in order to investigate more precisely any effect on learning. As effect sizes resulting from studies with larger samples contain fewer sampling errors and are therefore more precise and reliable estimators, effect sizes based on larger samples should be weighted more than those based on smaller samples. Hence, in meta-analysis, all data analysis is conducted with weighted effect sizes only. Each effect size was weighted by the inverse of its sampling error variance (Morris, 2008) and by an additional random variance component in order to take into account heterogeneity among the effect sizes (Hedges & Pigott, 2004).

Outlier Analysis

Extreme effect sizes are less representative for the research field and have a disproportionate influence on the computation of means and variances; therefore, they could produce misleading results. The distribution of effect sizes thus had to be examined in order to detect possible outliers (Hedges & Olkin, 1985). According to common procedures for handling outliers, effect sizes more than 1.5 times the interquartile range beyond the 25th or from the 75th percentile were adjusted to the respective inner fence value (Lipsey, 2009; Tukey, 1977). In this meta-analysis, no outliers were found beyond these bounds. In addition, we checked for influential data points that would have an excessive weight in the scope of the weighting. Across the 24 effect sizes for academic achievement, one would assume an average weight of 4.17. As the highest weight was 5.14, no highly influential data point was identified. The same applied to the 53 effect sizes for SRL (highest weight 2.54) and the 32 effect sizes for motivation (highest weight 3.89).

Dealing with Statistically Dependent Data

In meta-analysis, the unit of analysis is the primary research study. This becomes problematic when studies generate more than one effect size that measures the same construct. Whenever scores from multiple questionnaires or subscales are reported from the same sample, effect estimates from such clusters are statistically dependent, which can lead to risk of type I error inflation. To address these data structures, we used robust variance estimation (RVE, Hedges et al., 2010). RVE builds on the adjustment of standard errors and does not require precise information on the covariances between the effect sizes from the same clusters (Tipton & Pustejovsky, 2015). RVE meta-analysis on mean differences provides approximately correct confidence

intervals independently of the numbers of included clusters and estimates per cluster (Hedges et al., 2010).

In RVE, one assumes a certain correlation between the effect estimates within clusters that is the same for each cluster. If this expected correlation deviates from the true correlation, this does not result in bias, but only in a reduced efficiency. To calculate the weighted average effect sizes, we used RVE supposing a correlation of $\rho = .80$ between estimates within each cluster (Tanner-Smith & Tipton, 2014). We performed sensitivity tests with ρ values varying from $\rho = 0.0$ to $\rho = 0.9$, which confirmed that the results were robust to the choice of ρ. We used RVE for the overall meta-analysis that included correlational data as most studies provided multiple outcomes: on academic achievement, SRL, and motivation.

Theoretically, there is no need to analyze effects of monitoring tools for each of the three outcome variables separately as we do not have any hypotheses on the tools to work differently on the multiple outcomes. Yet, to test the robustness of our findings, we conducted sub-analyses for each outcome variable separately. For these sub-analyses, we could not compute RVE meta-regressions due to the smaller number of included studies. If the number of studies is not very large, confidence intervals from RVE meta-analyses tend to be biased when only a small number of clusters contribute multiple estimates but most clusters provide only one estimate (Hedges et al., 2010). Thus, for these sub-analyses, we use random-effects meta-regression for these outcomes, since most covariates in our analyses were categorical, producing a limited amount of between-cluster variation in the values of the covariates. To take into account the remaining statistical dependency of the data, multiple effect sizes resulting from one study and measuring the same construct were averaged so that only one effect size per study measuring one construct was included in the analysis. To this end, we conducted a fixed-effects meta-regression for each study (one per outcome) and entered the resulting average effect sizes per outcome into the three random-effect meta-regressions: one for academic achievement, one for SRL, and one for motivation (Rosenthal, 1991).

Heterogeneity Across Studies

To quantify the degree of heterogeneity of the effect sizes, I^2 was computed which determines the proportion of total variation in the estimates of the intervention effect that is due to heterogeneity between studies (Higgins & Thompson, 2002). I^2 was calculated using $I^2 = [(Q - \text{df}) = Q] \times 100\%$ (Tanner-Smith & Tipton, 2014). An I^2 of 0% would indicate that all variability in effect estimates is due to sampling error alone, and none is due to heterogeneity.

Analysis of Moderator Effects to Explain Heterogeneity

To explain the heterogeneity between the studies, studies were combined based on several factors possibly responsible for the effectiveness of the monitoring tools. We computed multiple meta-regressions to investigate the predictive value of the potential moderator variables. Multiple regression allows accounting for all potentially

important predictors in one model, i.e., to determine the relative influence of one or more predictor variables to the outcome. Using multiple meta-regression adjusts for multiple variables, and particularly for potential confounders, to better understand moderating effects (Tipton et al., 2019). Meta-regression differs from normal regression analyses mainly due to the weighting of effect sizes that are included as dependent variables into the regression function. Since the weights would be assumed to represent the numbers of subjects in a standard weighted regression procedure, significance testing would be based on incorrect assumptions regarding the sample size (Lipsey & Wilson, 2001). In meta-regression, the standard error for the regression slopes must therefore be corrected by the square root of the mean-square residual (Higgins & Thompson, 2004).

The analyses were conducted in Stata, version 16. For RVE meta-analyses, we employed the *robumeta* command (Hedges et al., 2010). For the fixed and random-effects meta-regressions, we used the *meta* commands (StataCorp, 2021).

Narrative Analyses of the Studies

Using the guidelines of Aveyard (2014), we started by developing codes in order to analyze the studies we found more in-depth. All papers were re-read thoroughly and this information was coded and compiled into a summary table (see Appendix Table 4): sample size; aim of the study; student age; duration in weeks; frequency of processing the monitoring tool; type of monitoring (cognitive, metacognitive, artefact collection); focus of the monitoring tool (learning content, learning behavior, both); format of the monitoring tool (open, closed, mixed); teacher feedback. In the next step, associated codes were organized into four themes: magnitude of effect sizes; timing of monitoring; active or passive engagement in goal setting; quality of feedback. These themes were used to present the information obtained from the primary studies in a narrative manner to provide more context that can be used to interpret the results of the quantitative meta-analysis.

Results

Our search identified 32 interventions, which resulted from 29 articles, and provided 109 independent effect sizes, including a total of 3492 participants. These intervention studies investigated the effectiveness of using a monitoring tool on academic achievement, strategy use, or motivation by using a longitudinal control group design.

Summary of Effect Sizes

Appendix Table 4 provides an overview of study characteristics and effect sizes per study. With regard to our research questions, 15 of the tools focused the learner's attention only on the learning content; nine only on the learning behavior. Eight of the tools focused on both (H2.1). Moreover, 18 of the studies used a tool to activate learners' metacognitive monitoring, while studies used a tool to stimulate cognitive

monitoring, and six studies used a tool to engage learners in collecting artifacts of their learning progress (H2.2). In half of the studies, participants received teacher feedback on their entries in the tool (H2.3). The average duration of the interventions was $M = 12.72$ weeks ($SD = 10.29$) (H2.4). The mean age of participants was $M = 17.17$ years ($SD = 1.87$; minimum = 9 years, maximum = 26 years), with one quarter of the studies being conducted with primary school students, one quarter with secondary school students, and the other half in higher education (H2.5). Among the 109 effect sizes retrieved were $k = 24$ effect sizes that assessed academic achievement, $k = 53$ effect sizes that measured SRL, and $k = 32$ effect sizes for motivation. All effect sizes that measured academic achievement were based on achievement tests.

The effect sizes that assessed SRL were based on commonly used self-report questionnaires which showed to be reliable and valid in studies that analyzed their internal consistency as well as their criterion validity (Klug et al., 2011). Only one study used a self-developed questionnaire (Dörrenbächer & Perels, 2016). Five of the studies used the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1993) or a translation of the MSLQ. Beside the MSLQ, the Reflective Thinking Scale (Kember et al., 2000), the Reporting Autonomous Studying Questionnaire (RAS; Elshout-Mohr et al., 2003), and the Student Learning Strategies Questionnaire (SLSQ; Abrami & Aslan, 2007) were applied. For most of the studies, Cronbach's alpha ranged within acceptable boundaries ($M = .74$; $SD = .11$).

Motivation was also assessed by means of self-report questionnaires. For three studies, the effect sizes for motivation could be retrieved from the motivational subscales of the MSLQ. In the other studies, the following questionnaires had been used: Academic Self-Regulation Questionnaire (SRQ; Ryan & Connell, 1989), Intrinsic Motivation Inventory (IMI; Deci & Ryan, 2006), Inventory for Learning Styles for Higher Education (ILS-HE; Vermunt, 1992), and the Self-Efficacy Scale (Schwarzer & Jerusalem, 1999). As with SRL, the reliabilities for the motivation scales were within an acceptable range ($M = .74$, $SD = .10$). Note that we only extracted the subscale that assessed motivation from questionnaires that assessed more than just motivation (for example, the subscale *situational interest enhancement* of the MSLQ [Pintrich et al., 1991] assessed in the study by Cazan, 2012). Subscales measuring SRL (for example, *time and study environmental management* in the MSLQ) were assigned to the outcome category SRL.

Examination of Publication Bias

In order to test for publication bias, we first constructed a funnel plot that plotted each study's standard error of effect sizes on the y-axis and the corresponding effect size on the *x*-axis (see Fig. 2). The resulting funnel plot does not indicate a publication bias of the type we predicted. There was a relation between the standard errors of the effect sizes and the effect sizes, but this was due to larger studies with more precise estimates reporting effect sizes that were closer to zero. This is most likely not due to publication bias, since that would mean that studies with null results

Fig. 2 Funnel Plot of Effect Sizes
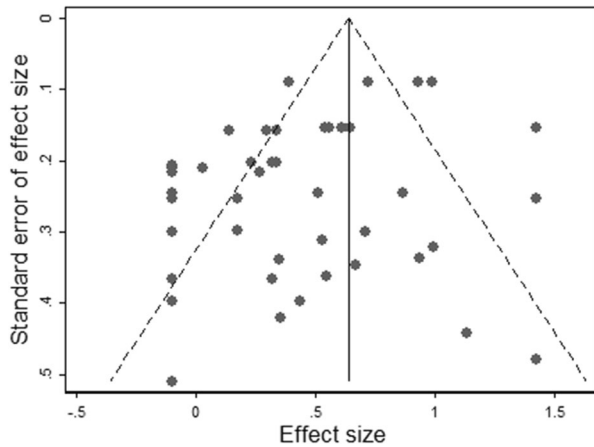
would be over-represented among studies, whereas studies with strong effects would have remained unpublished.

Second, we performed Egger's test of the intercept. The Egger test performs a linear regression of the effect estimates on their standard errors, weighting by 1/(variance of the effect estimate; Egger et al., 1997). The Egger test confirmed the asymmetry of the funnel plot. The intercept differed significantly from zero ($\beta = -1.68$; $p = 0.039$), 95% CI [-3.28; -0.085], and the negative value of this intercept indicates a negative relation between standard errors and effect sizes.

The pattern found in the funnel plot and in the Egger test is most likely due to heterogeneity among the reported studies (as will be further discussed in the following section, *Moderator Analyses*), with the larger studies having characteristics associated with smaller effect sizes.

Average Effect Sizes on Achievement, SRL, and Motivation

The effect sizes varied in magnitude. Figures 3, 4, and 5 represent forest plots of the 24 effect sizes for academic achievement, the 53 effect sizes for SRL, and the 32 effect sizes for motivation.

In line with our hypothesis H1, the empty random-effects models revealed statistically significant grand mean effect sizes for all three outcomes. The weighted mean effect size overall was moderate with $d = 0.27$ ($p = .0008$; 95% CI [.12, .42]; $k = 109$; $m = 31$), and for academic achievement showed a moderate effect of $d = 0.42$ ($p = .005$; 95% CI [.14, .70]; $k = 24$, $m = 18$). The effects were lower for SRL with $d = 0.19$ ($p = .02$; 95% CI [.04, .35]; $k = 53$, $m = 17$), and for motivation with $d = 0.17$ ($p = .04$; 95% CI [.01, .33]; $k = 32$, $m = 19$), but the difference between the effect sizes was not significant (see Table 2). Results of these null models suggest that learners who were exposed to an educational intervention based on the use of a monitoring tool showed significantly improved academic achievement and

Fig. 3 Forest plots of the effect sizes for academic achievement
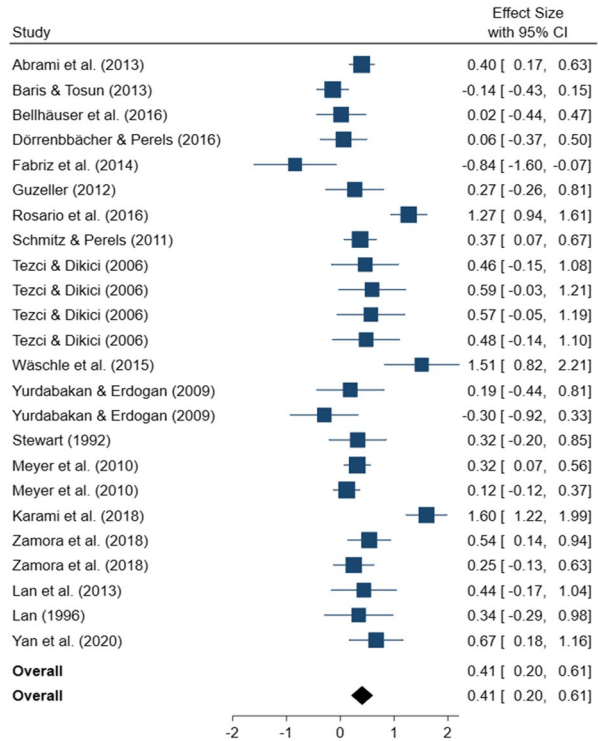
higher SRL and motivation when compared to learners who were not provided the intervention. Nevertheless, the random-effects null model yielded statistically significant heterogeneity. I^2 suggests that 70% of variability in point estimates is due to heterogeneity rather than sampling error, indicating that there is unaccounted variability in the individual effect sizes used to calculate the overall unbiased effect. These findings revealed that there was substantial heterogeneity in between-study differences based on study characteristics. In order to capture these differences in more detail and to decipher the reasons for this heterogeneity, moderator analyses were conducted with respect to the focus of the monitoring tool (learning content, learning behavior, or both), the type of monitoring stimulated by the tool (cognitive monitoring, metacognitive monitoring, or collection of artifacts), external teacher feedback on the tool entries (yes, no), the duration of the intervention, and the age of the participants.

Moderator Analyses

Overall Meta-analysis

Moderator analyses were performed with meta-analytic regressions to explain heterogeneity between studies. In the first step, we conducted an overall meta-analysis

Fig. 4 Forest plots of the effect sizes for self-regulated learning

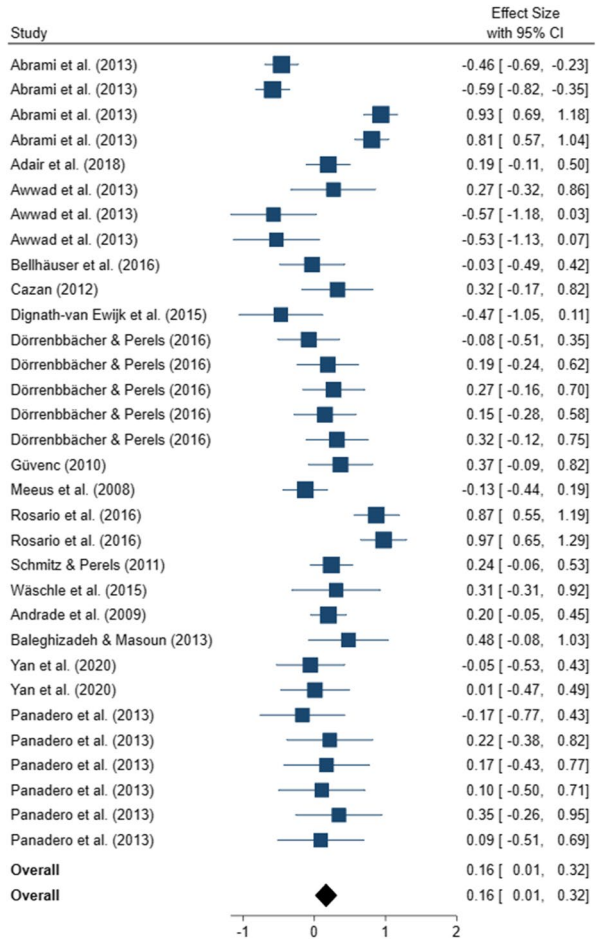| Study	Effect Size with 95% CI
Abrami et al. (2013)	0.27 [0.03, 0.50]
Abrami et al. (2013)	0.07 [-0.16, 0.30]
Abrami et al. (2013)	0.03 [-0.20, 0.26]
Abrami et al. (2013)	0.25 [0.02, 0.48]
Abrami et al. (2013)	0.24 [0.01, 0.47]
Abrami et al. (2013)	0.31 [0.07, 0.54]
Adair et al. (2018)	-0.03 [-0.33, 0.28]
Awwad et al. (2013)	0.44 [-0.16, 1.04]
Awwad et al. (2013)	0.10 [-0.49, 0.69]
Awwad et al. (2013)	-0.45 [-1.05, 0.14]
Awwad et al. (2013)	0.25 [-0.35, 0.84]
Bellhäuser et al. (2016)	0.00 [-0.46, 0.46]
Cazan (2012)	0.60 [0.09, 1.10]
Cazan (2012)	0.60 [0.10, 1.11]
Dignath-van Ewijk et al. (2015)	0.42 [-0.16, 1.00]
Dörrenbbächer & Perels (2016)	-0.28 [-0.71, 0.15]
Dörrenbbächer & Perels (2016)	0.07 [-0.37, 0.50]
Dörrenbbächer & Perels (2016)	0.01 [-0.42, 0.45]
Dörrenbbächer & Perels (2016)	-0.21 [-0.65, 0.22]
Dörrenbbächer & Perels (2016)	0.35 [-0.08, 0.79]
Dörrenbbächer & Perels (2016)	0.40 [-0.03, 0.84]
Dörrenbbächer & Perels (2016)	0.26 [-0.18, 0.69]
Fabriz et al. (2014)	-0.04 [-0.77, 0.69]
Fabriz et al. (2014)	-0.12 [-0.74, 0.50]
Greenwood (2010)	-0.38 [-0.71, -0.05]
Greenwood (2010)	0.15 [-0.18, 0.47]
Greenwood (2010)	-0.06 [-0.38, 0.27]
Greenwood (2010)	0.12 [-0.21, 0.45]
Greenwood (2010)	-0.28 [-0.60, 0.05]
Güvenc (2010)	0.62 [0.15, 1.08]
Meeus et al. (2008)	-0.24 [-0.55, 0.07]
Meeus et al. (2008)	0.26 [-0.06, 0.57]
Rosario et al. (2016)	0.68 [0.37, 0.99]
Schmitz & Perels (2011)	0.36 [0.06, 0.65]
Meyer et al. (2010)	0.27 [0.03, 0.52]
Meyer et al. (2010)	0.42 [0.18, 0.67]
Meyer et al. (2010)	0.42 [0.17, 0.66]
Meyer et al. (2010)	0.47 [0.23, 0.72]
Meyer et al. (2010)	-0.20 [-0.44, 0.05]
Altiok et al. (2019)	0.95 [0.43, 1.47]
Altiok et al. (2019)	0.70 [0.19, 1.20]
Altiok et al. (2019)	0.56 [0.06, 1.06]
Altiok et al. (2019)	0.70 [0.19, 1.20]
Altiok et al. (2019)	0.36 [-0.14, 0.85]
Altiok et al. (2019)	0.59 [0.08, 1.09]
Altiok et al. (2019)	0.91 [0.40, 1.43]
Altiok et al. (2019)	1.02 [0.50, 1.54]
Yan et al. (2020)	-0.05 [-0.53, 0.43]
Yan et al. (2020)	-0.12 [-0.60, 0.36]
Panadero et al. (2013)	-0.89 [-1.51, -0.26]
Panadero et al. (2013)	-0.38 [-0.99, 0.22]
Panadero et al. (2013)	0.75 [0.13, 1.37]
Panadero et al. (2013)	-0.10 [-0.70, 0.50]
Overall	0.21 [0.11, 0.30]
Overall	0.21 [0.11, 0.30]

Fig. 5 Forest plots of the effect sizes for motivation

Study	Effect Size with 95% CI
Abrami et al. (2013)	-0.46 [-0.69, -0.23]
Abrami et al. (2013)	-0.59 [-0.82, -0.35]
Abrami et al. (2013)	0.93 [0.69, 1.18]
Abrami et al. (2013)	0.81 [0.57, 1.04]
Adair et al. (2018)	0.19 [-0.11, 0.50]
Awwad et al. (2013)	0.27 [-0.32, 0.86]
Awwad et al. (2013)	-0.57 [-1.18, 0.03]
Awwad et al. (2013)	-0.53 [-1.13, 0.07]
Bellhäuser et al. (2016)	-0.03 [-0.49, 0.42]
Cazan (2012)	0.32 [-0.17, 0.82]
Dignath-van Ewijk et al. (2015)	-0.47 [-1.05, 0.11]
Dörrenbbächer & Perels (2016)	-0.08 [-0.51, 0.35]
Dörrenbbächer & Perels (2016)	0.19 [-0.24, 0.62]
Dörrenbbächer & Perels (2016)	0.27 [-0.16, 0.70]
Dörrenbbächer & Perels (2016)	0.15 [-0.28, 0.58]
Dörrenbbächer & Perels (2016)	0.32 [-0.12, 0.75]
Güvenc (2010)	0.37 [-0.09, 0.82]
Meeus et al. (2008)	-0.13 [-0.44, 0.19]
Rosario et al. (2016)	0.87 [0.55, 1.19]
Rosario et al. (2016)	0.97 [0.65, 1.29]
Schmitz & Perels (2011)	0.24 [-0.06, 0.53]
Wäschle et al. (2015)	0.31 [-0.31, 0.92]
Andrade et al. (2009)	0.20 [-0.05, 0.45]
Baleghizadeh & Masoun (2013)	0.48 [-0.08, 1.03]
Yan et al. (2020)	-0.05 [-0.53, 0.43]
Yan et al. (2020)	0.01 [-0.47, 0.49]
Panadero et al. (2013)	-0.17 [-0.77, 0.43]
Panadero et al. (2013)	0.22 [-0.38, 0.82]
Panadero et al. (2013)	0.17 [-0.43, 0.77]
Panadero et al. (2013)	0.10 [-0.50, 0.71]
Panadero et al. (2013)	0.35 [-0.26, 0.95]
Panadero et al. (2013)	0.09 [-0.51, 0.69]
Overall	0.16 [0.01, 0.32]
Overall	0.16 [0.01, 0.32]

combining the 109 effect sizes. To account for variability across outcome variables, we included the three outcome variables as predictors in our model. As shown in Table 3, the results did not vary as a function of outcome variable. As is the case in multiple regressions, the inserted dummy variables are compared to a constant. In our analyses, the constant was academic achievement, metacognitive monitoring,

Table 2 Weighted mean effect sizes

	k	Number of studies m	Weighted mean ES	$-95\%CI$	$+95\%CI$	SE	Z	P
Overall	109	32	.27	.12	.42	.07	3.52	.0008
Academic achievement	24	18	.42	.14	.70	.13	1.33	.005
SRL	53	19	.19	.04	.34	.07	2.79	.01
Motivation	32	17	.17	.01	.33	.07	1.88	.03

Table 3 Moderator analyses for the overall meta-analysis

Moderator	Beta	SE	−95%CI	+95%CI	df	p
Self-regulated learning	−0.08	0.15	−0.40	0.23	17.68	.60
Motivation	−0.06	0.14	−0.34	0.22	18.79	.66
Focus on learning behavior	−0.40	0.11	−0.64	−0.15	12.52	.005
Focus on learning content	−0.03	0.18	−0.42	0.36	10.07	.85
Theory self-assessment	−0.44	0.18	−0.83	−0.04	10.67	.03
Theory portfolio	−0.06	0.25	−0.64	0.52	7.50	.81
Teacher feedback	0.32	0.23	−0.19	0.82	11.21	.20
Duration	−0.03	0.01	−0.05	−0.01	13.29	.007
Age	−0.006	0.02	−0.05	0.04	9.36	.77

Constant is academic achievement, theory self-regulated learning, focus on learning behavior and learning content, no teacher feedback

focus on both learning content and learning behavior, and no teacher feedback. This overall meta-analysis revealed that the effect varied as a function of the focus of the tool (H2.1). For studies that used tools to engage learners in metacognitive monitoring and did not include teacher feedback, effects were moderately larger when the tool focused on both learning content and learning behavior than only on learning behavior ($B = −0.40$, $SE = 0.11$, $p = .005$), but not larger than for tools that focused on learning content only. Moreover, as expected, the effect differed depending on the type of monitoring stimulated with the tool (H2.2) with higher effects when tools simulated metacognitive monitoring compared to cognitive monitoring ($B = −0.44$, $SE = 0.18$, $p = .03$). No significant difference was found for stimulating the collection of artefacts by means of portfolio. On a descriptive level, effect sizes were moderately higher when learners obtained external feedback by the teacher (H2.3) than in studies without feedback, but this difference was not significant ($B = 0.32$, $SE = 0.23$, $p = .19$. Furthermore, we found that the effect size declined with an increasing duration of the intervention (H2.4), measured in number of weeks ($B = −0.03$, $SE = 0.007$, $p < .001$), indicating that longer intervention is not beneficial. Finally, our analysis did not reveal a moderation effect for age (H2.5), ($B = −0.006$, $SE = 0.02$, $p = .77$). The results of the moderator analyses are displayed in Table 3.

Separate Meta-analyses per Outcome

In the second step, separate meta-analyses were conducted for the different outcome categories to confirm the robustness of the results. In the first step, we computed a separate meta-analysis for each study with multiple outcomes for achievement to compute one estimate for achievement per study in which we meta-analytically summarized the estimates from each given study on each given outcome to a single estimate by running within-study fixed-effects meta-analyses. The same was done for studies with multiple outcomes measuring SRL and motivation. Next, we conducted a random-effects meta-regression combining the 19 average effects of academic

achievement, another meta-regression for the 19 average effects measuring SRL, and another meta-regression for the 17 average effects of motivation. Not all moderators were significant in these three meta-analyses; most likely due to substantially reduced power. Yet, altogether the meta-regressions confirm the pattern found in the overall meta-analysis.

Academic Achievement. For academic achievement, most of the findings from the overall meta-analysis could be confirmed on a descriptive level. Again, the constant was metacognitive monitoring, focus on both learning content and learning behavior, and no teacher feedback. Compared to this constant, studies using tools for both learning content and learning behavior produced higher effect sizes than tools for learning behavior only, ($B = -0.998$, $SE = 0.54$, $p = .019$). On a descriptive level, metacognitive monitoring was more beneficial than cognitive monitoring ($B = -1.11$, $SE = 0.48$, $p = .064$). However, this result was no longer significant at the 5% level. Moreover, no significant moderator effect was found for teacher feedback ($B = 0.31$, $SE = 0.38$, $p = .42$). Like in the overall analyses, interventions with shorter duration were more effective for achievement ($B = -0.06$, $SE = 0.03$, $p = .03$. As expected, no moderator affect for age (H6) could be found ($B = -0.02$, $SE = 0.04$, $p = .62$).

SRL. The meta-analysis for SRL confirmed the overall analysis, but partly only on a descriptive level. Effect sizes were significantly higher when the tool engaged learners in metacognitive monitoring compared to cognitive monitoring ($B = -0.67$, $SE = 0.22$, $p = .003$). The moderator effects for the focus of the tool (H2.1) was not found to be significant ($B = -0.27$, $SE = 0.21$, $p = .196$). Like in the achievement meta-analysis, teacher feedback (H2.3) was not fount to moderate the effect. The moderating effect of the duration of the intervention (H2.4) just missed the 5% significance level ($B = -0.02$, $SE = 0.01$, $p = .068$). Finally, no age differences were found.

Motivation. Interestingly, this meta-regression did not confirm H2.1 with a significant advantage of tools focusing on both learning content and learning behavior over tools that focus on learning behavior only ($B = -0.003$, $SE = 0.21$, $p = .987$), but we found higher effects on motivation when the tools focused only on learning content instead of both content and learning behavior. However, this effect did not reach the 5% significance level ($B = 0.34$, $SE = 0.19$, $p = .07$). The meta-regression for motivation indicated higher effect sizes for tools that stimulated metacognitive monitoring compared to cognitive monitoring ($B = -0.37$, $SE = 0.20$, $p = .06$), but this result narrowly failed the 5% significance level. Again, no difference was found for portfolio tools. On a descriptive level, effect sizes were higher when teachers provided feedback, but this effect was not significant at the 5% level ($B = 0.38$, $SE = 0.28$, $p = .167$). As in academic achievement, the duration of the intervention moderated the effects on motivation (H2.4) with shorter interventions being more effective ($B = -0.04$, $SE = 0.02$, $p = .03$). With regard to effects on motivation, the age was found as a moderator (H2.5), indicating that younger learners benefitted more from the intervention ($B = -0.03$, $SE = 0.01$, $p = .025$).

Overall, the results of the separate meta-regressions confirmed the pattern of moderator variables found in the overall meta-analysis, although some moderator effects failed to reach the significance level, probably due to the reduced number of effect sizes leading to less power. The main difference found between outcome variables was that tools which focused the learners' attention to both the learning

content and their learning behavior were more effective to enhance achievement and SRL than tools that addressed only learning behavior, and tools focusing on the learning content only yielded higher effects on motivation. In addition, with regard to motivation effects, studies were more effective when they focused on younger learners, and—at least on a descriptive level—when teachers provided feedback.

Narrative Analyses of the Studies

In order to analyze the studies in more depth and to be able to make more precise statements about the effectiveness of the moderators, we additionally subjected the studies to a narrative analysis. To this end, we coded contextual information (student age; duration in weeks; frequency of processing the monitoring tool; external feedback by teachers or peers), and content-related information about the intervention (type of monitoring; focus of the monitoring tool; format of the monitoring tool). Out of the 31 interventions, 14 took place in primary or secondary school; the remaining interventions were conducted in the context of higher education. Based on the coded information (see Appendix Table 4), we derived the four following themes: (a) magnitude of effect sizes; (b) timing of monitoring; (c) active or passive engagement in goal setting; (d) quality of feedback.

The Most Effective Studies

In the first step, we grouped the studies according to the magnitude of their effects on academic achievement to get an overview of associations between potential moderators and effect sizes. Eighteen of the studies provided data to compute effect sizes for academic achievement. According to Hattie (2013), an educational intervention must achieve at least an effect size of $d = 0.40$ to be practically relevant. This is based on the fact that $d = 0.40$ is the average effect found in his meta-synthesis across all educational interventions. In our meta-analysis, seven studies had an effect on achievement of at least $d = 0.40$ or greater. Two of these studies were conducted with students in higher education (Karami et al., 2018; Lan et al., 1993); the remaining studies were carried out with students between 9 and 14 years old. In all of these studies, the monitoring was not only retrospective, but students had to work on the monitoring tool during task execution. Whereas two of these studies used a monitoring tool that consisted of collecting artefacts of learning (Karami et al., 2018; Tezci & Dikici, 2006), the other five studies used a tool that encouraged metacognitive monitoring. In most of these highly effective studies, learners were not only encouraged to monitor and evaluate their learning progress, but also to reflect on appropriate strategies for improving their own learning. For example, Wäschle et al. (2015) provided the learners with prompts to stimulate the use of cognitive and metacognitive strategies when writing a learning journal entry. Likewise, Yan et al. (2020) encouraged the learners to identify their strengths and weaknesses based on the quality of their possible strategies for improvement. Abrami et al. (2013) stimulated learners to reflect on their achievement and also on their strategies, and to use these reflections to regulate their learning goals for the next work. In summary, the studies

that produced the highest effects differed from other studies in that (a) learners were required to complete the monitoring tool *during* task completion, and (b) control strategies were encouraged in addition to monitoring.

Student Age

Sorting the study results in terms of learner age reveals that, except for the studies by Karami et al. (2018) and Lan et al. (1993) which took place in higher education, all studies with practically significant achievement effects were conducted in the school context. However, the picture is different with regard to effectiveness in increasing SRL and motivation: only in the study by Rosario et al. (2017) we found high effects in the school context. In the context of higher education, practically significant effects on SRL (Altiok et al., 2019; Cazan, 2012; Dignath et al., 2015) and motivation (Baleghizadeh & Masoun, 2013) were found in studies where achievement was not examined as an outcome. With regard to age effects, it is difficult to draw conclusions as the findings are inconclusive. This is in agreement with the results of the meta-analysis.

Intervention Duration

An ambiguous picture emerges when looking at the studies according to the intervention duration. Here, high-performance effects are found in studies with very short (Wäschle et al., 2015; Yan et al., 2020), medium (Lan et al., 1993; Rosario et al., 2017), and very long intervention durations (Abrami et al., 2013; Karami et al., 2018). The situation is different with regard to effects on SRL and on motivation. Here, the highest effects are found in studies with a medium intervention duration.

Timing of the Monitoring

There was no clear pattern as to whether learners should make a prospective assessment before they start learning (i.e., before task execution) in addition to a retrospective assessment (i.e., after task execution). While two of the 11 studies that required learners to provide prospective assessments resulted in practically significant effects (Abrami et al., 2013; Yan et al., 2020), most studies that resulted in very high effects did not require learners to provide prospective assessments about their learning (e.g., Rosario et al., 2017; Wäschle et al., 2015). Very clear, however, are the results for the timing of monitoring: practically significant and high effects on achievement are only found in studies in which learners have to work with the monitoring tool during task processing. However, high effects for SRL (Altiok et al., 2019; Cazan, 2012; Güvenc, 2010) and for motivation (Baleghizadeh & Masoun, 2013) are shown in individual studies in which learners did not need to process the monitoring tool during task execution, but only afterwards. The results suggest that there are positive effects on achievement development when learners work on the monitoring tool during task processing. However, this does not seem to have such clear effects on the development of SRL and motivation.

Active or Passive Engagement in Setting Standards

In six studies, learners were encouraged to set their own goals or standards. One of these studies was conducted in primary school (Andrade et al., 2009), and two of the studies were conducted in secondary school (Güzeller, 2012; Tezci & Dikici, 2006); the other half was carried out in higher education. Only one of these studies resulted in significant performance effects (Tezci & Dikici, 2006) and another in relevant effects for SRL (Cazan, 2012). Otherwise, practically significant or high effects were found across all outcome variables only in studies in which learners were not encouraged to formulate standards or goals themselves. Thus, no recommendations can be derived from the results at this time that learners should derive monitoring goals or standards on their own.

Quality of Teacher Feedback

In 11 of the studies, learners received feedback on their entries in the monitoring tool. Seven of these studies were conducted in the school context; three of them in primary school (Abrami et al., 2013; Andrade et al., 2009; Stewart, 1992). This did not show higher effects for a particular age group of learners. Of these studies, three studies had a short duration (< 10 weeks), and four studies were long interventions of 20–28 weeks (Abrami et al., 2013; Andrade et al., 2009; Greenwood, 2010; Stewart, 1992). Of these feedback studies, one resulted in very high effects on achievement (Karami et al., 2018), and two others resulted in moderate effects on achievement (Abrami et al., 2013; Tezci & Dikici, 2006). Furthermore, the intervention by Altiok et al. (2019) produced medium effects on SRL, and Baleghizadeh and Masoun (2013) showed medium effects on motivation.

We then investigated whether the learners were not only passive recipients of teacher feedback but also had to give peer feedback themselves and were thus more actively involved in the feedback process. In four of the feedback studies, learners were not only passive recipients of feedback, but were also able to provide peer feedback to their classmates themselves. Two of these studies resulted in significant effects on achievement (Abrami et al., 2013) and SRL (Altiok et al., 2019), respectively, while the other two studies resulted in null effects. Thus, additional peer feedback did not increase the effectiveness of teacher feedback in these studies.

Next, we examined which of the studies allowed learners to revise their work in response to teacher feedback. In five of the feedback studies, learners had the opportunity to revise, and all three feedback studies that resulted in high effects on achievement also allowed learners to do so. However, the two studies that resulted in substantial effects on SRL and motivation, respectively, did not provide this opportunity. Teacher feedback that also provides the opportunity to act on that feedback, then, appears to be particularly effective for achievement. Interestingly, the narrative analyses thus show that teacher feedback can also have positive effects on achievement (and not only on motivation, as indicated in the quantitative analyses) if the feedback includes the possibility to revise one's work on the basis of the feedback.

To sum up, the most effective studies required students to use the monitoring tool *during* task execution, whereas most other studies had students process the tool before or after learning. Moreover, most of these studies encouraged metacognitive

monitoring, and stimulated the learners to reflect on their strategy use in addition to reflecting on their performance. Even though teacher feedback was not a significant moderator in the meta-analysis, it could still contribute to effectiveness beyond motivation if learners were also given the opportunity to revise their work based on the feedback; i.e., to process the feedback directly.

Discussion

To become lifelong learners in an advanced technological society, students need to develop monitoring skills. Unfortunately, a majority of students are not strong self-regulated learners yet (European Council, 2006) and are poor at monitoring their own learning (Brown & Harris, 2013; Greene & Azevedo, 2007). These students' progress and achievement are at risk as SRL skills are required for successful learning at school and in post-secondary education (Dresel et al., 2015). This meta-analysis presents two major key findings related to the effectiveness of monitoring interventions. First, the weighted mean effect size of the 32 studies revealed positive effects ($d = 0.27$, 95% CI [.12, .42]) overall, as well as on academic achievement ($d = 0.42$, 95% CI [.14, .70]), on SRL ($d = 0.19$, 95% CI [.04, .34]) and on motivation ($d = 0.17$, 95% CI [.01, .33]) in favor of monitoring interventions compared with a control group that received no intervention. In practice, this means that learners who are encouraged to engage in some form of monitoring show improved performance, strategy use, and motivation with respect to their learning. As such, integrated monitoring has a positive impact on learner performance and self-regulated learning skills.

Second, there was substantial heterogeneity in the weighted mean effect sizes between studies ($I^2 = 70\%$), warranting the need to account for the impact of moderator variables on the mean effect size. In view of the substantial heterogeneity of study effects, five moderator variables were considered in this meta-analysis: the focus of the monitoring tool, the type of monitoring stimulated by the tool, external teacher feedback, the duration of the intervention, and the age of the participants. Results from multiple random-effects meta-regression models showed that three of the five moderators significantly moderated the overall mean effect size. The findings from this meta-analysis show that there is a variety of tools that foster monitoring and that improve learners' achievement and motivation. The available evidence on such tools, however, is broader than the studies included in our meta-analysis as we did not include studies without a control group or without a pretest measure or studies with a single-subject design, qualitative, or correlational studies.

The Reactivity Effect of Monitoring Tools

Our most important finding is that the use of a monitoring tool improved academic achievement substantially, suggesting that using a tool to encourage active participation in the monitoring process has positive effects on student learning. Thus, these findings support the assumption of a reactivity effect (Zimmerman, 2002) on achievement. In general, the present meta-analysis is consistent with the findings

of previous meta-analytic studies on monitoring interventions. Brown and Harris (2013) synthesized the literature on self-assessment interventions for children from kindergarten through grade 12, and found a similar effect on academic achievement (median: $d = 0.40$). Likewise, Guzman et al. (2018) investigated the effectiveness of self-monitoring on achievement and found an effect size Tau-$U = 0.79$ which corresponds to $d = 0.47$. Sanchez et al. (2017) also found a moderate effect of self-grading interventions on achievement ($g = 0.34$), and, comparably, Yan et al. (2021) found a moderate effect of $g = 0.45$ for self-assessment interventions on achievement. Thus, our findings concur with former evidence from systematic reviews in the field, which have been yielding moderate effects around 0.40. Hattie (2012) argues that educational interventions must achieve at least an effect greater than $d = 0.40$ to be considered practically relevant, as this was the average effect size of all the interventions he studied in his meta-synthesis.

With regard to effects on SRL, Harkin et al. (2016) examined the effectiveness of interventions on monitoring one's goal and also found an effect of $d = 0.40$. A comparable effect was found by Gutierrez de Blume 2022 in a meta-analysis of learning strategy intervention on learners' monitoring accuracy ($g = 0.56$). Panadero et al. (2017), however, found lower effects of self-assessment interventions on SRL ($d = 0.23$), which are more similar to the findings of the meta-analysis presented here. Concerning motivation, we can only compare our results to Panadero et al.'s (2017) meta-analysis that examined the effects of self-assessment on self-efficacy. Contrary to our findings, however, that revealed only low effects on motivation ($d = 0.17$), they found high effect sizes for self-efficacy ($d = 0.73$; Panadero et al., 2017). These differences in findings may be because the current meta-analysis captured motivation at a broader level (e.g., motivation to learn), whereas Panadero et al. (2017) focused specifically on self-efficacy.

At large, these results are promising as they show that journaling, logging, self-assessment, and other tools which are used to promote monitoring—and which have been widely used in educational practice—are effective and can be recommended. Nevertheless, the effectiveness of monitoring tools differs substantially according to several characteristics of the tool, the implementation, and the learner. In the following, we will derive more specific recommendations from the findings of our moderator analyses and the narrative analyses of the studies.

Center the Attention on the Learning Content and the Learning Behavior

Concerning the focus of the monitoring tool, studies that used tools focusing on both—monitoring of the learning content and of the learning behavior—produced significantly higher effects on achievement than studies that used a tool stimulating monitoring of the learning behavior. In our meta-analysis, all the studies with tools that solely prompted monitoring of learning behavior only, used highly structured questionnaires as monitoring tools (e.g., Bellhäuser et al., 2016; Dignath et al., 2015; Schmitz & Perels, 2011). These learning questionnaires were supposed to prompt learners to monitor and reflect their learning behavior, but the items used in these questionnaires are usually independent of the learning content and thus rather prompt monitoring of learning strategies than monitoring of understanding (example item: "Today I used aids (internet, encyclopedias, …) for my homework."; Dignath et al., 2015).

Contrary to these questionnaire-like learning diaries, learning journals with an open answering format prompt learners to monitor their understanding of content. Either questions address the content (e.g., "Which examples can you think of that illustrate, confirm or conflict with the learning contents?"; Wäschle et al., 2015) or the learner's understanding of the content (e.g., "Which main points do you now understand, and which haven't you understood?"; Wäschle et al., 2015). Having this closer look at the tools shows that tools with a stronger focus on the learning content and the learning behavior are addressing both the cognitive monitoring of one's understanding and the metacognitive monitoring of one's use of learning strategies, whereas tools that focus on the learning behavior only tend to skip the monitoring of understanding. These findings are in line with the COPES model that assumes that metacognitive monitoring of the self-regulated learning behavior cannot take place without cognitive monitoring of a learner's understanding of the learning content (Winne, 2004).

Furthermore, our findings showed that tools that only focused on monitoring of the learning content even produced highest effects on motivational outcomes. In general, research shows that the effort needed to complete such tools on a regular basis has motivational costs (Nückles et al., 2021). For example, Nückles et al. (2010) found that students who received cognitive and metacognitive prompts did not increase their use of cognitive strategies at the end of the term, while students with open learning journals intuitively increased their use. Moreover, those students who used more cognitive strategies at the end of the term produced higher learning outcomes. Providing students with cognitive and metacognitive prompts shortly before taking an exam apparently prohibits students from increasing their use of cognitive strategies, probably as a result of increasing cognitive load (Nückles et al., 2010). This could explain why students might reap long-term benefits from an increase in self-regulation strategies, as in the beginning the additional effort might be detrimental to their learning of academic content and to their motivation.

The results thus suggest that it is more effective if a tool stimulates the monitoring of the learning content in any case, and only additionally a monitoring of the learning behavior. An attempt should be made to keep learners motivated if monitoring both areas is too much at once. One possibility could be to familiarize learners with the monitoring of the learning content first, and only add the monitoring of the learning behavior in a second step. One method could be to practice monitoring both areas separately and very concretely, for example with the help of worked examples.

The Differing Impetus of Tools for Cognitive and Metacognitive Monitoring and Portfolio Tools

Effect sizes were larger when the study employed a tool to stimulate metacognitive monitoring compared to tools that activate cognitive monitoring. One possible reason could be that additionally focusing one's attention on one's learning and understanding—as in metacognitive monitoring—may be more proactive for learning and achievement than just cognitively evaluating one's understanding and performance against a certain standard without further embedding it in its own learning behavior. Moreover, metacognitive monitoring is probably more comprehensive than cognitive monitoring

since metacognitive monitoring cannot do without cognitive monitoring and therefore integrates both. This is partly because monitoring focuses more on the process and, thus, leads to direct implications for regulation processes, whereas cognitive monitoring focuses more on the product, which does not necessarily result in implications for regulation. This is also confirmed by the narrative analyses of the studies: the most effective studies complemented monitoring with a reflection on their own learning and the appropriate strategies, thus stimulating not only monitoring but also control processes. This result raises the question of whether an instrument to promote monitoring should not generally be combined with the promotion of control measures.

Cognitive monitoring can provide learners with transparency on the teachers' expectations (Panadero & Jonsson, 2013). Technically, learners can use a tool for cognitive monitoring also to plan, monitor, and evaluate their learning (Jonsson, 2014); however, some studies showed that learners perceived higher stress when being confronted with cognitive monitoring and rather used avoidance-oriented learning strategies than learning-oriented SRL strategies (e.g., Panadero & Romero, 2014). Whereas cognitive monitoring focuses more on the expectations that learners have to meet, and often remain summative instead of formative (Panadero & Jonsson, 2020), metacognitive monitoring focuses on students' learning experiences, which are critical for effective self-regulation during learning (Griffin et al., 2013). Tools for metacognitive monitoring that emerge from SRL theory usually strive to provide learners with a holistic framework to monitor and also to adapt their learning strategies (Cleary et al., 2008). Thus, rather than concentrating on the current state, metacognitive monitoring tools based in SRL theory aim to engage learners in regulation processes to improve the current state.

Does Teacher Feedback Boost the Reactivity Effect?

Contrary to former research (Graham et al., 2015; Hattie, 2016), we did not find a significant impact of teacher feedback. Although teacher feedback was a positive moderator of the effect on motivation, this effect was not significant on the 5% level. Feedback is considered a primary component of formative assessment (Black & Wiliam, 1998; Hattie & Timperley, 2007); however, it is not a homogenous concept but can differ in many ways, including the feedback's agents, its content and its implementation (Panadero & Lipnevich, 2022). First, however, it is an important finding that feedback may moderate effects on motivation, as we found at least at the descriptive level. For example, it is motivating for learners if something happens with their efforts when working on the monitoring tools. Beyond this appreciation, however, feedback can also be motivating when it enables learners to improve their learning by providing them with concrete ideas about how they can modify their learning.

In our meta-analysis, 11 studies reported that participants received teacher feedback. And the majority of the tools used in these studies focused on the learning content only, and only very few of these studies were grounded in SRL theory (c.f., Abrami et al., 2013; Lan, 1996; Meyer et al., 2010). It is therefore possible to assume that the applied feedback focused only on the learning content as well, thereby

providing rather more evaluative (as in the case of self-assessment) than proactive feedback as suggested by SRL theory. And whereas also content-related feedback is known to differ in its effectiveness, depending on accompanying information such as grades or praise (Lipnevich & Smith, 2009), it takes more to engage learners in SRL. They need scaffolding to manage the transfer from cognitive and metacognitive monitoring to the next planning phase. The pure act of monitoring one's learning is not yet self-regulated. The learner still has to work with the outcome of the monitoring process and has to transform the conclusions of the self-assessment into new plans for the next learning phase (Zimmerman, 2000).

Hattie and Timperley (2007) presented a model on feedback that includes the steps of *feed up* for the planning phase ("Where am I going?"), *feedback* for the monitoring phase ("How am I going?"), and *feed-forward* for the transition from the evaluation phase to the next planning phase ("Where to next?"). Completely self-regulated learners would be able to answer these questions on their own when working with a monitoring tool. Most students, however, still need some guidance and scaffolding in order to answer such questions effectively—this can be provided through teacher feedback. When working with tools for cognitive and metacognitive monitoring, this translates into providing students with the outcome of their reflection and assistance in choosing strategies for how to proceed. Feedback should not only refer to the task and process level, but also address the SRL and the self-level (Chou & Zou, 2020; Yan, 2020). As such tools help to regularly and formatively assess progress, teachers can use them to provide more powerful feedback than just the simple snapshot of summative assessment (Hattie, 2012). Thus, the absence of a significant moderating effect of feedback in our meta-analysis could be related to its quality and may not generally exclude the boosting effect of feedback.

Although teacher feedback was not a significant moderator of effects on achievement in this meta-analysis, it may still contribute to effectiveness beyond motivation, as our narrative analyses suggest. If learners are also given the opportunity to revise their work based on teacher feedback, that is, to process the feedback directly, then the studies have shown practically relevant effects. In the sense of Hattie and Timperley (2007), feedback should not only contain feedback about strengths and weaknesses of one's own work (feedback), but also point out possible approaches to transform one's own weaknesses into strengths (feedforward).

More is Not Always More

Contrary to our expectation, the effectiveness of SRL training did not increase with the length of the intervention; we even found the opposite effect: shorter intervention led to larger effects. Earlier findings on the length of the intervention have been inconsistent. Whereas Dignath and Büttner (2008) found the effectiveness of SRL training to increase with the duration of the training, De Boer et al. (2018) did not find that duration moderated training effect. However, to our knowledge, no former meta-analysis in this field found an opposite effect of duration.

This could in part be due to motivation problems. Working with the same tool for a long time can be perceived as boring and the learner might suffer from a lack of motivation due to the additional workload (see, for example, Dignath et al., 2015). In addition,

working with a monitoring tool is not the same as a strategy training. Learners are prompted to monitor, but they are not (necessarily) provided with control strategies to react to the monitoring result. For this reason, using the tool in the long run may not provide additional opportunities for practice, and, thus, does not have much additional effect.

Monitoring Tools Seem to Work for All

The findings of our meta-analysis suggest that monitoring tools are equally effective for foster SRL and achievement among learners of all age groups. This result is in line with other meta-analytic evidence, suggesting that formative and self-assessment are equally effective among younger and older learners (see Dent & Koenka, 2016; Kingston & Nash, 2011; Panadero et al., 2017). Nevertheless, our qualitative review of the studies suggests that only two studies carried out in the context of higher education led to substantial effects on achievement, whereas six studies in primary and secondary school led to effect sizes higher than $d = 0.40$, and four more studies yielded effects above $d = 0.30$. This finding suggests that it might be more beneficial to practice monitoring at an early stage with the help of a monitoring tool and to embed it into regular lessons. This result is strengthened by the moderator analyses' finding that effects on motivation are higher for younger learners than for older ones.

In contrast, the narrative analyses showed higher effects for SRL and motivation in older learners. On the whole, then, our qualitative and quantitative findings suggest that monitoring tools are effective for all target groups, but that there are likely age-specific effects for different outcomes. For example, since the effects on motivation seem to diminish with age, it seems important to include elements that keep learners motivated, especially with older students. The results are encouraging because they suggest that it is not too late to practice monitoring; that is, older students are not yet too fixed in their learning behavior to benefit from such an intervention. On the other hand, they also show that already very young learners can learn monitoring and they can thus build up a strategy repertoire at an early stage from which they can benefit throughout their school years.

Meta-analysis Identifies Specific Research Needs

Meta-analyses not only serve to investigate effectiveness and to offer recommendations, but also to provide a comprehensive overview of the research field and to identify specific research needs. First of all, the small number of studies meeting the eligibility criteria of this meta-analysis is surprising and disappointing. Given the replication crisis in psychology, the limited number of studies we found for such an important topic is concerning. One reason for the small number of studies is that we only included studies that had a pre-post control group design. To test causality, experimental or at least quasi-experimental designs are required (Schneider, 2007). Although random assignment of students to intervention conditions is not always possible, at least a quasi-experimental control group design with pre- and posttest allows testing the effects of natural development (Grant et al., 2013). However, many studies in this field do not use a control group or a pretest and thus do not meet methodological standards. In our meta-analysis, we chose to include only studies that were experimental or quasi-experimental; however, this was at

the expense of power, as it allowed us to find only a small number of studies. As our results show, the research field urgently needs more (quasi-)experimental studies to replicate findings on the effectiveness of monitoring tools and to clarify open questions.

Another important issue is the paucity of studies examining monitoring at early school age. Although we did not use an age restriction in our research, we did not find any studies conducted with learners younger than 9 years old. Therefore, we must limit our generalizations to monitoring practices from later primary school age onward. But this focus on learners of older elementary school age and beyond may not be coincidental. Similar age limitations in the studies found were also evident in Sanchez et al.'s (2017) meta-analysis of the effectiveness of self-grading interventions. It would be possible that teachers believe that children 8 years of age or younger have not yet developed the metacognitive skills necessary for effective monitoring. In fact, there is a lack of systematic research on the age at which students can learn monitoring with the help of such tools. Future studies should systematically examine how the effectiveness of monitoring tools differs for beginning school children compared to more experienced learners.

Another research brief arises from the operationalization of SRL in most studies in this meta-analysis. The valid measurement of SRL represents an unsolved problem that has been increasingly raised in the last decades. The bulk of the assessment of SRL is based on self-report questionnaires (Roth et al., 2016). The results of our meta-analysis show this quite clearly: self-report was used almost exclusively to measure SRL. However, research has shown that self-report questionnaires suffer from a lack of validity (Veenman, 2011). Metacognitive processes are not observable in themselves, which makes retrospective self-report difficult. If SRL was not validly captured, the effects on SRL in this meta-analysis may be correspondingly inaccurate. Accordingly, these studies may not have captured whether the use of monitoring tools really had an effect on SRL.

In addition, research has shown that SRL is situational rather than constant across situations and subjects (Winne, 2004), meaning learners cannot easily answer items about strategy use in general (Veenman & van Cleef, 2019). This raises the question of whether the effect sizes for SRL found in our review represent the type of SRL that should be addressed by means of the tools that foster self-monitoring (state measures) or whether the outcome variables of the primary studies assess something other (trait measures) than the intended SRL competence (Winne, 2010). The questionnaires used in the primary studies of this meta-analysis are not likely to capture potential effects if they assess different strategies than the ones that are prompted by the tools. If the measurement error in such self-report measures of SRL is classical, this would bias estimated correlations between SRL measurements and the use of monitoring tools toward zero. This is illustrated in the findings of the meta-analysis by Dent and Koenka (2016), which revealed significant differences between the type of SRL measure: self-report measures led to the lowest effect sizes ($d = .17$ for self-report questionnaires and $d = .16$ for self-report interviews) compared to speech ($d = 48$) or behavior during task ($d = .37$). Comparably, Panadero et al. (2017) found lower effects of self-assessment on SRL when SRL was measured with self-report strategy questionnaires ($d = .23$) than for SRL measured qualitatively ($d = .43$). There is a lack of research using appropriate measurement instruments that can capture the construct of SRL, in particular with regard to the use of metacognitive strategies (planning, self-monitoring, and self-reflection) that are central to the principle of self-monitoring. More research

is needed to test the effects of tools on SRL in a more sophisticated way to identify the mechanism by which the use of monitoring tools enhances academic achievement. In particular, in studies where learners are working on a monitoring tool, it would be easy to capture SRL as a process measure. In addition, there are technology-based ways to capture SRL if the monitoring tool is implemented digitally.

Another research issue that emerged in our meta-analysis is the lack of precision in the presentation of the studies. The narrative analyses of the included studies show that the concrete intervention elements are described too vaguely, which makes it difficult to systematically investigate their mechanisms of action. For example, information about when learners processed the monitoring tool, about the standards against which learners compared their learning progress, and about what teacher feedback exactly entailed and how it was implemented (e.g., what the content of the feedback was and what happened to the feedback as it progressed) was sometimes so vaguely described that it was difficult to code. Feedback that is not heard or acted upon by learners is unlikely to have an impact, and if the feedback is procedural or too fact-based, it is probably not going to show any effects. But it was not possible to reconstruct this detailed information from many of the studies in our meta-analysis, and hardly any studies systematically examined the effectiveness of different types of feedback or implementation of feedback. In order to make statements about the effectiveness of feedback in the context of monitoring tools, more in-depth research is needed that is precise enough to specifically test the elements of feedback and its implementation. It is essential that researchers document their approach and the implemented interventions very precisely.

The last thing that stands out in the studies is that the monitoring tools were implemented in authentic learning situations, mostly even in the natural classroom; yet, teachers were only marginally included in the studies. Monitoring tools not only offer learners the opportunity to reflect on their learning progress. They can also give teachers exciting insights into their students' thinking and learning, which they could use to provide targeted feedback to guide their students as they develop into self-regulated learners.

Limitations

Our synthesis is subject to the typical limitations arising from the nature of meta-analyses. In particular, our analysis was constrained by the degree of completeness of reporting in the identified primary studies. For example, instructional procedures of the implementation of the tools have been described carefully in some studies, but not in others. Because of limited information about the details of implementation, we could, for example, not investigate whether the specific type of prompt used in the tool, or whether the completion of the tool was graded by the teacher, would moderate the impact of the treatment. As these details of implementation are important to know in order to provide clear guidelines for educational practice, these research questions should be addressed systematically in future primary research. Questions regarding statistical power limited the study of many moderators, such that the number of moderators was small or unevenly distributed among groups of moderator variables. Moreover, this meta-analysis was limited to quasi-experimental or experimental studies that compared the effects of

a treatment against a control group. Focusing on these types of studies does not disregard the contribution of other types of research, such as qualitative studies or single-subject design studies. Furthermore, as we restricted our review to studies that investigated the effects of monitoring tools, which were implemented for longer periods, we did not include studies that investigated the impact of such tools in a singular session such as a laboratory experiment (e.g., Panadero & Romero., 2014). Thus, we cannot draw any conclusions about the effectiveness of the unique use of monitoring tools and consequent moderators. With regard to the stability of the effects, we could not perform a follow-up meta-analysis to test long-term effects because only two studies (Güzeller, 2012; Wäschle et al., 2015) provided follow-up data to compute effect sizes. As in all effectiveness research on academic achievement, the goal is to improve learning sustainably. Since the theoretical principle of tools, such as learning journals, is based on the improvement of SRL (in terms of monitoring and control), one would assume that students would need some time to change their SRL before effects in academic achievement become visible. A delayed effect of several weeks or months can therefore be assumed, which makes follow-up measures necessary. Finally, the studies included in this review did not all apply the same measure for academic achievement. To consider this issue, we calculated effect sizes from holistic measures (i.e., overall scores) whenever possible or transformed more specific measures into one average effect size per study (by computing a meta-analysis across the outcome measures for academic achievement within one study). Nevertheless, there was variability concerning the academic content of the measure (e.g., reading, writing, or mathematics), the attributes assessed (e.g., within the field of writing), the scale points of the measure, and the operationalization of these points. This variability should be acknowledged when interpreting the findings of this review.

Summary and Conclusion

Altogether, this meta-analysis provides the first quantitative summary of effectiveness research of the use of monitoring tools on learner outcomes. As an implication for a theoretical model of SRL, the findings of our meta-analysis align with the notion that such tools can serve to stimulate learning and have a positive effect on academic achievement and motivation (Schmitz & Perels, 2011). However, as our results show, not every tool is equally effective. Hence, in the studies we integrated, particularly high effects are found when the tool (1) not only addresses learning behavior but also learning behavior and content in combination, (2) stimulates not only cognitive monitoring but also metacognitive monitoring, and (3) is not only complemented by teacher feedback but learners also have the opportunity to implement this feedback directly. Finally, our findings suggest that using such tools works for all age groups of learners equally well. However, alongside learners' age and expectations in different educational contexts, their expertise in terms of prior knowledge may also moderate the effectiveness of these tools; earlier research has provided evidence for an expertise reversal effect when using learning journals (Nückles et al., 2010). Future research could examine how to adapt the use of monitoring tools to the expertise level of the learner. Finally, our findings lead to a plea to use more sophisticated and valid study designs and measurement to assess the effects of these tools on learners' SRL.

Appendix

Table 4 Overview of studies included in the meta-analysis

Author (year of publication)	N	N_{effect}	Aim of the study	Theoretical background	Focus	Format	Age	Duration in weeks	Frequency	External feedback
Abrami et al. (2013)	319	11	ACH, SRL, MOT	SRL	Both	Open or mixed	11	28	Once per week	Yes
Adair et al. (2018)	181	2	SRL, MOT	PORT	Learning content	Open or mixed	20	12	Twice per week	No
Altok et al. (2019)	70	8	SRL	SRL	Learning content	Open	20	7	Once per week	Yes
Andrade et al. (2009)	268	1	MOT	SRL	Learning content	Closed	10	20	n.s.	Yes
Awwad et al. (2013)	49	7	SRL, MOT	SRL	Learning content	Closed	22	16	n.s.	No
Baleghizadeh & Masoun (2013)	57	1	MOT	ASSESS	Learning content	Closed	26	7	Less than once per week	Yes
Baris & Tosun (2013)	199	1	ACH	PORT	Learning content	Open or mixed	16	18	n.s.	No
Bellhäuser et al. (2016)	83	3	ACH, SRL, MOT	SRL	Learning behavior	Closed	20	4	Daily	No
Cazan (2012)	117	3	SRL, MOT	SRL	Both	Open or mixed	20	7	Once per week	No
Dignath-van Ewijk et al. (2015)	44	2	SRL, MOT	SRL	Learning behavior	Closed	24	13	Once per week	No
Dörrenbächer & Perels (2016)	91	13	ACH, SRL	SRL	Learning behavior	Closed	23	7	Daily	No
Fabriz et al. (2014) (1) EG₁ (EFL) (2) EG₂ (Psychology)	(1) 32 (2) 45	(1) 2 (2) 1	ACH, SRL	SRL	Learning behavior	Closed	23	14	Twice per week	No
Greenwood (2010)	158	5	SRL	SRL	Learning behavior	Open	15	20	Once per week	Yes
Güvenç (2010)	84	2	SRL, MOT	SRL	Learning behavior	Open or mixed	19	10	Once per week	No
Güzeller (2021)	60	1	ACH	ASSESS	Learning content	Open or mixed	16	5	Twice per week	Yes
Karami et al. (2018)	151	1	ACH	PORT	Learning content	Open	21	16	Once per week	Yes
Lan (1996)	44	1	ACH	SRL	Both	Closed	20	15	Twice per week	No
Lan et al. (1993)	48	1	ACH	SRL	Both	Closed	20	15	Twice per week	No
Meeus et al. (2008)	174	3	SRL, MOT	PORT	Learning behavior	Open or mixed	20	16	n.s.	No

Table 4 (continued)

Author (year of publication)	N	N_{effect}	Aim of the study	Theoretical background	Focus	Format	Age	Duration in weeks	Frequency	External feedback
Meyer et al. (2010)	296	7	ACH, SRL	SRL	Both	Closed	11	28	Once per week	Yes
Panadero et al. (2013) (1) EG₁ (rubrics) (2) EG₂ (scripts)	69	(1) 3 (2) 7	SRL, MOT	ASSESS	Learning content	Closed	20	10	Once per week	No
Rosário et al. (2017)	182	4	ACH, SRL, MOT	SRL	Learning content	Open	12	12	Once per week	No
Schmitz & Perels (2011)	195	3	ACH, SRL, MOT	SRL	Learning behavior	Closed	13	7	Daily	No
Stewart (1992)	62	1	ACH	SRL	Learning content	Open	9	24	Three times per week	Yes
Tezci & Dikici (2006)	52	4	ACH	PORT	Learning content	Closed	14	16	Twice per week	Yes
Wäschle et al. (2015)	46	2	ACH, MOT	SRL	Both	Open or mixed	13	3	Once per week	No
Yan et al. (2020)	74	5	ACH, SRL, MOT	ASSESS	Both	Closed	12	5	N.s.	No
Yurdabakan & Tolga (2009)	44	2	ACH	PORT	Learning content	Closed	15	12	Twice per week	Yes
Zamora et al. (2018) (1) EG₁ (2) EG₂	198	(1) 1 (2) 1	ACH	ASSESS	(1) Both (2) Learning content	Closed	13	2	Once per week	No

MOT, motivation; *ACH*, learning achievement; *ASSESS*, self-assessment; *PORT*, portfolio

Declarations

References

References marked with an asterisk were included in the meta-analysis.

Abrami, P. C., & Aslan, O. (2007). *The student learning strategies questionnaire*. Centre for the Study of Learning & Performance, Concordia University.

*Abrami, P. C., Venkatesh, V., Meyer, E. J., & Wade, C. A. (2013). Using electronic portfolios to foster literacy and self-regulated learning skills in elementary students. *Journal of Educational Psychology*, *105*(4), 1188–1209. https://doi.org/10.1037/a0032448

*Adair, D., Jaeger, M., & Price, O. M. (2018). Promoting active learning when teaching introductory statistics and probability using a portfolio curriculum approach. *International Journal of Higher Education*, *7*(2), 175–188. https://doi.org/10.5430/ijhe.v7n2p175

*Altiok, S., Başer, Z., & Yükseltürk, E. (2019). Enhancing metacognitive awareness of undergraduates through using an e-educational video environment. *Computers & Education*, *139,* 129–145. https://doi.org/10.1016/j.compedu.2019.05.010

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2008.00118.x

Andrade, H. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research, 102*(4), 287–302. https://doi.org/10.3200/JOER.102.4.287-302

*Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice*, 48(1), 12–19. https://doi.org/10.1080/00405840802577544

Aveyard, H. (2014). *Doing a literature review in health and social care: A practical guide*. Open University Press.

*Awwad, F. M. A., Nofal, M. B., & Salti, N. S. (2013). The impact of electronic portfolio on developing reflective thinking and self-directed learning readiness. *Cypriot Journal of Educational Sciences*, *8*(1), 78–104.

*Baleghizadeh, S., & Masoun, A. (2013). The effect of self-assessment on EFL learners' self-efficacy. *TESL Canada Journal*, *31*(1), 42–58.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58.

*Baris, F. M., & Tosun, N. (2013). Can social networks and e-portfolio be used together for enhancing learning effects and attitudes? *The Turkish Online Journal of Educational Technology*, *12*, 31–62.

Beckers, J., Dolmans, D. H., & van Merriënboer, J. J. (2019). Perflect: Design and evaluation of an electronic development portfolio aimed at supporting self-directed learning. *TechTrends, 63*(4), 420–427.

*Bellhäuser, H., Lösch, T., Winter, C., & Schmitz, B. (2016). Applying a web-based training to foster self-regulated learning — Effects of an intervention for large numbers of participants. *The Internet and Higher Education*, *31*, 87–100. https://doi.org/10.1016/j.iheduc.2016.07.002

Berthold, K., Nückles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction, 17*(5), 564–577. https://doi.org/10.1016/j.learninstruc.2007.09.007

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics, 57*(2), 213–227. https://doi.org/10.1023/B:EDUC.0000049293.55249.d4

Brown, G. T., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367–393). Sage.

Brown, G. T., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education, 28*(7), 968–978.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(5), 245–281. https://doi.org/10.2307/1170684

Casadevante, C., Ortner, T., Romero, M., Hernández, J. M., & Santacreu, J. (2021). How to foster learning by limiting the response speed. *Current Psychology*. Advance online publication. https://doi.org/10.1007/s12144-021-01784-7

*Cazan, A.-M. (2012). Enhancing self-regulated learning by learning journals. *Procedia - Social and Behavioral Sciences*, *33*, 413–417. https://doi.org/10.1016/j.sbspro.2012.01.154

Chang, C.-C. (2008). Enhancing self-perceived effects using web-based portfolio assessment. *Computers in Human Behavior, 24*(4), 1753–1771. https://doi.org/10.1016/j.chb.2007.07.005

Chang, C.-C., Liang, C., & Chen, Y.-H. (2013). Is learner self-assessment reliable and valid in a web-based portfolio environment for high school students? *Computers & Education, 60*(1), 325–334. https://doi.org/10.1016/j.compedu.2012.05.012

Chou, C.-Y., & Zou, N.-B. (2020). An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00233-y

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205–249. https://doi.org/10.1007/s10648-011-9191-6

Clarke, S., Timperley, H., & Hattie, J. (2003). *Unlocking formative assessment: Practical strategies for enhancing students' learning in the primary and intermediate classroom*. Hodder Moa Beckett.

Cleary, T. J., Platten, P., & Nelson, A. (2008). Effectiveness of the self-regulation empowerment program with urban high school students. *Journal of Advanced Academics, 20*(1), 70–107. https://doi.org/10.4219/jaa-2008-866

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen Goodman, M. (1998). *The effect of journal writing on the reading comprehension, study habits and attitude of underprepared college students*. Temple University.

Costa Ferreira, P., Veiga Simão, A. M., & Lopes da Silva, A. (2015). Does training in how to regulate one's learning affect how students report self-regulated learning in diary tasks? *Metacognition and Learning, 10*(2), 199–230. https://doi.org/10.1007/s11409-014-9121-3

De Boer, H., Donker, A. S., Kostons, D. D., & van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review, 24*, 98–115. https://doi.org/10.1016/j.edurev.2018.03.002

Deci, E. L., & Ryan, R. M. (2006). Intrinsic motivation inventory. http://www.selfdeterminationtheory.org

Delclos, V. R., & Harrington, C. (1991). Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology, 83*(1), 35–42. https://doi.org/10.1037/0022-0663.83.1.35

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

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.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? *Educational Research Review, 3*(2), 101–129. https://doi.org/10.1016/j.edurev.2008.02.003

*Dignath-van Ewijk, C., Fabriz, S., & Büttner, G. (2015). Fostering self-regulated learning among students by means of an electronic learning diary: A training experiment. *Journal of Cognitive Education and Psychology*, *14*(1), 77–97. https://doi.org/10.1891/1945-8959.14.1.77

Donker, A. S., De Boer, H., Kostons, D., Dignath-van Ewijk, C. D., & van der Werf, M. P. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review, 11*(1), –26. https://doi.org/10.1016/j.edurev.2013.11.002

*Dörrenbächer, L., & Perels, F. (2016). More is more? Evaluation of interventions to foster self-regulated learning in college. *International Journal of Educational Research*, *78*, 50–65. https://doi.org/10.1016/j.ijer.2016.05.010

Dresel, M., Schmitz, B., Schober, B., Spiel, C., Ziegler, A., Engelschalk, T., et al. (2015). Competencies for successful self-regulated learning in higher education: Structural model and indications drawn from expert interviews. *Studies in Higher Education, 40*(3), 454–470. https://doi.org/10.1080/03075079.2015.1004236

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Elshout-Mohr, M., Meijer, J., van Daalen-Kapteijns, M. M., & Meeus, W. (2003). *A self-report inventory for metacognition related to academic tasks*. University of Amsterdam, SCO Kohnstamm Instituut.

European Council. (2006). Recommendation of the European Parliament and the Council on key competences for lifelong learning. *Official Journal of the European Union, 49*, 10–18 https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:en:PDF

Fabriz, S., Dignath, C., Poarch, G., & Büttner, G. (2007). Förderung von Selbstregulationskompetenz mit standardisierten Lerntagebüchern [Promotion of self-regulation competence with standardized learning diaries]. In M. Krämer, S. Preiser, & K. Brusdeylins (Eds.), *Psychologiedidaktik und Evaluation VI* (pp. 95–104). V & R unipress.

*Fabriz, S., Dignath-van Ewijk, C., Poarch, G., & Büttner, G. (2014). Fostering self-monitoring of university students by means of a standardized learning journal–A longitudinal study with process analyses. *European Journal of Psychology of Education, 29*(2), 239–255. https://doi.org/10.1007/s10212-013-0196-z

Fazio, L. L. (2001). The effect of corrections and commentaries on the journal writing accuracy of minority- and majority-language students. *Journal of Second Language Writing, 10*(4), 235–249. https://doi.org/10.1016/S1060-3743(01)00042-X

Glaser, C., & Brunstein, J. C. (2007). Improving fourth-grade students' composition skills: Effects of strategy instruction and self-regulation procedures. *Journal of Educational Psychology, 99*(2), 297–310. https://doi.org/10.1037/0022-0663.99.2.297

Glogger, I., Schwonke, R., Holzäpfel, L., Nückles, M., & Renkl, A. (2012). Learning strategies assessed by journal writing: Prediction of learning outcomes by quantity, quality, and combinations of learning strategies. *Journal of Educational Psychology, 104*(2), 452–468. https://doi.org/10.1037/a0026683

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Grant, S. P., Mayo-Wilson, E., Melendez-Torres, G. J., & Montgomery, P. (2013). Reporting quality of social and psychological intervention trials: A systematic review of reporting guidelines and trial publications. *PLoS One, 8*(5), e65442. https://doi.org/10.1371/journal.pone.0065442

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334–372. https://doi.org/10.3102/003465430303953

*Greenwood, J. C. (2010). *The effect of reflective portfolio use on student self-regulation skills in science (Publication No. 10258795)*. , Western Connecticut State University, ProQuest Dissertations Publishing.

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In *International handbook of metacognition and learning technologies* (pp. 19–34). Springer.

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

Güven, G., Sülün, Y., & Çam, A. (2014). The examination of elementary preservice teachers' reflective diaries and epistemological beliefs in science laboratory. *Teaching in Higher Education, 19*(8), 895–907. https://doi.org/10.1080/13562517.2014.934350

*Güvenç, H. (2010). The effects of cooperative learning and learning journals on teacher candidates' self-regulated learning. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1477–1487.

*Güzeller, C. O. (2012). The effect of web-based portfolio use on academic achievement and retention. *Asia Pacific Education Review*, *13*(3), 457–464.

Guzman, G., Goldberg, T. S., & Swanson, H. L. (2018). A meta-analysis of self-monitoring on reading performance of K–12 students. *School Psychology Quarterly, 33*(1), 160–168. https://doi.org/10.1037/spq0000199

Harkin, B., Webb, T. L., Chang, B. P. I., Prestwich, A., Conner, M., Kellar, I., Benn, Y., & Sheeran, P. (2016). Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. *Psychological Bulletin, 142*(2), 198. https://doi.org/10.1037/bul0000025

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge. https://doi.org/10.4324/9780203181522

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

Hattie, J. (2016). Know thy impact. In M. Scherer (Ed.), *EL Essentials on formative assessment. Readings from educational leadership* (pp. 36–45). ASCD.

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*(2), 99–136. https://doi.org/10.3102/00346543066002099

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445. https://doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T., & Thompson, S. G. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*(11), 1559–1573. https://doi.org/10.1002/sim.1187

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23*(11), 1663–1682. https://doi.org/10.1002/sim.1752

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840–852. https://doi.org/10.1080/02602938.2013.875117

*Karami, S., Sadighi, F., Bagheri, M. S., & Riasati, M. J. (2018). The potential impact of the application of electronic portfolio on Iranian EFL learners' writing performance seeking their gender role. *Cogent Social Sciences*, *4*(1). https://doi.org/10.1080/23311886.2018.1562509

Kember, D., Leung, D. Y. P., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., & Wong, M. (2000). Development of a questionnaire to measure the level of reflective thinking. *Assessment & Evaluation in Higher Education, 25*(4), 381–395. https://doi.org/10.1080/713611442

Khodadady, E., & Khodabakhshzade, H. (2012). The effect of portfolio and self assessment on writing ability and autonomy. *Journal of Language Teaching and Research, 3*(3), 518–524. https://doi.org/10.4304/jltr.3.3.518-524

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klug, J., Ogrin, S., Keller, S., Ihringer, A., & Schmitz, B. (2011). A plea for self-regulated learning as a process: Modelling, measuring and intervening. *Psychological Test and Assessment Modeling, 53*, 51–72.

*Lan, W. Y. (1996). The effects of self-monitoring on students' course performance, use of learning strategies, attitude, self-judgment ability, and knowledge representation. *The Journal of Experimental Education*, *64*(2), 101–115. https://doi.org/10.1080/00220973.1996.9943798

*Lan, W. Y., Bradley, L., & Parr, G. (1993). The effects of a self-monitoring process on college students' learning in an introductory statistics course. *The Journal of Experimental Education*, *62*(1), 26–40. https://doi.org/10.1080/00220973.1993.9943829

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*(4), 319–333. https://doi.org/10.1037/a0017841

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders, 4*(2), 124–147. https://doi.org/10.1080/15564880802612573

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis. Applied social research methods series: V. 49*. Sage Publications.

Matas, C. P., & Allan, C. (2004). Using learning portfolios to develop generic skills with on-line adult students. *Australian Journal of Adult Learning, 44*, 7–26.

McAuley, L., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet, 356*(9237), 1228–1231. https://doi.org/10.1016/S0140-6736(00)02786-0

*Meeus, W., van Petegem, P., & Meijer, J. (2008). Stimulating independent learning: A quasi-experimental study on portfolio. *Educational Studies*, *34*(5), 469–481. https://doi.org/10.1080/03055690802288478

*Meyer, E. J., Abrami, P. C., Wade, C. A., Aslan, O., & Deault, L. (2010). Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL. *Computers & Education*, *55*(1), 84–91. https://doi.org/10.1016/j.compedu.2009.12.005

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Myers, M., & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology, 70*(5), 680–690. https://doi.org/10.1037/0022-0663.70.5.680

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

Nückles, M. (2021). Investigating visual perception in teaching and learning with advanced eye-tracking methodologies: Rewards and challenges of an innovative research paradigm. *Educational Psychology Review, 33*(1), 149–167.

Nückles, M., Hübner, S., Dümer, S., & Renkl, A. (2010). Expertise reversal effects in writing-to-learn. *Instructional Science, 38*(3), 237–258. https://doi.org/10.1007/s11251-009-9106-9

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271. https://doi.org/10.1016/j.learninstruc.2008.05.002

Olina, Z., & Sullivan, H. J. (2002). Effects of classroom evaluation strategies on student achievement and attitudes. *Educational Technology Research and Development, 50*(3), 61–75. https://doi.org/10.1007/BF02505025

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Panadero, E., Fernández-Ruiz, J., & Sánchez-Iglesias, I. (2020). Secondary education students' self-assessment: The effects of feedback, subject matter, year level, and gender. *Assessment in Education: Principles, Policy & Practice, 27*(6), 607–634. https://doi.org/10.1080/0969594X.2020.1835823

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*. https://doi.org/10.1016/j.edurev.2020.100329

Panadero, E., Andrade, H., & Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher, 45*(1), 13–31.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*(422). https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Brown, G. T., & 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

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133–148. https://doi.org/10.1080/0969594X.2013.877872

*Panadero, E., Alonso-Tapia, J., & Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies in Educational Evaluation*, *39*(3), 125–132. https://doi.org/10.1016/j.stueduc.2013.04.001

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes portfolio a portfolio? *Educational Leadership, 48*(5), 60–63.

Perry, N. E., & Rahim, A. (2011). Studying self-regulated learning in classrooms. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 122–136). Routledge.

Pintrich, P., Smith, D., García, T., & McKeachie, W. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. University of Michigan.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*(1), 19–33. https://doi.org/10.1207/s15326985ep2501_3

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J., & 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

Roebers, C. M., Röthlisberger, M., Cimeli, P., Michel, E., & Neuenschwander, R. (2011). School enrolment and executive functioning: A longitudinal perspective on developmental changes, the influence of learning context, and the prediction of preacademic skills. *European Journal of Developmental Psychology, 8*, 526–540. https://doi.org/10.1080/17405629.2011.571841

*Rosário, P., Högemann, J., Núñez, J. C., Vallejo, G., Cunha, J., Oliveira, V., Fuentes, S., & Rodrigues, C. (2017). Writing week-journals to improve the writing quality of fourth-graders' compositions. *Reading and Writing*, *30*(5), 1009–1032. https://doi.org/10.1007/s11145-016-9710-4

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Sage.

Roth, A., Ogrin, S., & Schmitz, B. (2016). Assessing self-regulated learning in higher education: A systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability, 3*, 225–250. https://doi.org/10.1007/s11092-015-9229-2

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*(5), 749–761. https://doi.org/10.1037/0022-3514.57.5.749

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*, 1049–1066. https://psycnet.apa.org/doi/10.1037/edu0000190

*Schmitz, B., & Perels, F. (2011). Self-monitoring of self-regulation during math homework behaviour using standardized diaries. *Metacognition and Learning*, *6*(3), 255–273. https://doi.org/10.1007/s11409-011-9076-6

Schmitz, B., & Schmidt, J. K. M. (2011). Assessing self-regulated learning using diary measures with university students. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 265–280). Routledge.

Schneider, S. L. (2007). Experimental and quasi-experimental designs in behavioral research: On context, crud, and convergence. In W. Outhwaite, & S. P. Turner (Eds.). *The SAGE handbook of social science methodology* (pp. 174–189). https://doi.org/10.4135/9781848607958

Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer-und Schülermerkmalen*. In *Dokumentation der Psychometrischen Verfahren im Rahmen der wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education, 9*(2), 169–191.

StataCorp. (2021). *Stata: Release 16. Statistical Software*. StataCorp LLC.

*Stewart, C. B. (1992). *Journal writing in mathematics classrooms: A practical inquiry* (Publication No. 9311458 [Dissertation Memphis State University]. ProQuest Dissertations Publishing.

Stoeger, H., Fleischmann, S., & Obergriesser, S. (2015). Self-regulated learning (SRL) and the gifted learner in primary school: The theoretical basis and empirical findings on a research program dedicated to ensuring that all students learn to regulate their own learning. *Asia Pacific Education Review, 16*(2), 257–267. https://doi.org/10.1007/s12564-015-9376-7

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. http://www.jstor.org/stable/23363562

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*(1), 13–30. https://doi.org/10.1002/jrsm.1091

*Tezci, E., & Dikici, A. (2006). The effects of digital portfolio assessment process on students' writing and drawing performances. *The Turkish Online Journal of Educational Technology*, *5*(2), 46–55.

Tican, C., & Taspinar, M. (2015). The effects of reflective thinking-based teaching activities on pre-service teachers' reflective thinking skills, critical thinking skills, democratic attitudes, and academic achievement. *Anthropologist, 20*(1, 2), 111–120. https://doi.org/10.1080/09720073.2015.11891730

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods, 10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Todd, R. W. (2002). Using self-assessment for evaluation. The. *Forum, 40*, 16–19.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning, 6*(2), 205–211. https://doi.org/10.1007/s11409-011-9080-x

Veenman, M. V., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM, 51*(4), 691–701.

Vermunt, J. D. (1992). Leerstijlen en sturen van leerprocessen in het hoger onderwijs – naar procesgerichte instructie in zelfstandig denken [Learning styles and the direction of learning processes in higher education - towards process-oriented instruction in independent thinking]. Swets & Zeitlinger.

Wang, Y., & Sperling, R. A. (2020). Characteristics of effective self-regulated learning interventions in mathematics classrooms: A systematic review. *Frontiers in Education*, 1–17. https://doi.org/10.3389/feduc.2020.00058

*Wäschle, K., Gebhard, A., Oberbusch, E.-M., & Nückles, M. (2015). Journal writing in science: Effects on comprehension, interest, and critical reflection. *Journal of Writing Research*, *7*(1), 41–64. https://doi.org/10.17239/jowr-2015.07.01.03

Webber, J., Scheuermann, B., McCall, C., & Coleman, M. (1993). Research on self-monitoring as a behavior management technique in special education classrooms: A descriptive review. *Remedial and Special Education, 14*(2), 38–56. https://doi.org/10.1177/074193259301400206

Whitebread, D., Bingham, S., Grau, V., Pino Pasternak, D., & Sangster, C. (2007). Development of metacognition and self-regulated learning in young children: Role of collaborative and peer-assisted

learning. *Journal of Cognitive Education and Psychology, 6*(3), 433–455. https://doi.org/10.1891/194589507787382043

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Lawrence Erlbaum Associates Publishers.

Winne, P. H. (2004). Comments on motivation in real-life, dynamic, and interactive learning environments: Theoretical and methodological challenges when researching motivation in context. *European Psychologist, 9*(4), 257–263. https://doi.org/10.1027/1016-9040.9.4.257

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*(4), 267–276. https://doi.org/10.1080/00461520.2010.517150

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). Erlbaum Associates.

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G. J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human Computer Interaction, 35*(4–5), 356–373. https://doi.org/10.1080/10447318.2018.1543084

*Yan, Z., Chiu, M. M., & Ko, P. Y. (2020). Effects of self-assessment diaries on academic achievement, self-regulation, and motivation. *Assessment in Education: Principles, Policy & Practice*, *27*(5), 562–583. https://doi.org/10.1080/0969594X.2020.1827221

Yan, Z., Wang, X., Boud, D., & Lao, H. (2021). The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. *Assessment & Evaluation in Higher Education*, 1–15. https://doi.org/10.1080/02602938.2021.2012644

Yan, Z., & Brown, G. T. (2017). A cyclical self-assessment process: towards a model of how students engage in self-assessment. Assessment & Evaluation in Higher Education, 42(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., & Carless, D. (2021). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.2001431

*Yurdabakan, I., & Tolga, E. (2009). The effects of portfolio assessments on reading, listening and writing skills of secondary school prep class students. *The Journal of International Social Research*, 2, 526–538.

*Zamora, Á., Suárez, J. M., & Ardura, D. (2018). Error detection and self-assessment as mechanisms to promote self-regulation of learning among secondary education students. *The Journal of Educational Research, 111*(2), 175–185. https://doi.org/10.1080/00220671.2016.1225657

Ziegler, N. A. (2014). Fostering self-regulated learning through the European Language Portfolio: An embedded mixed methods study. *The Modern Language Journal, 98*(4), 921–936. https://doi.org/10.1111/modl.12147

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. https://doi.org/10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeider (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.

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

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. 311–328). Routledge.

Authors and Affiliations

Charlotte Dignath[1] · **Reyn van Ewijk[2]** · **Franziska Perels[3]** · **Sabine Fabriz[4]**

✉ Charlotte Dignath
charlotte.dignath@tu-dortmund.de

Reyn van Ewijk
vanewijk@uni-mainz.de

Franziska Perels
f.perels@mx.uni-saarland.de

Sabine Fabriz
fabriz@psych.uni-frankfurt.de

[1] Center for Research on Education and School Development (IFS), TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany

[2] Chair of Statistics and Econometrics, University of Mainz, Mainz, Germany

[3] University of the Saarland, Saarbrücken, Germany

[4] Institute of Educational Psychology, Goethe University, Frankfurt am Main, Germany