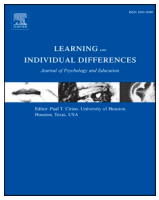
Learning and Individual Differences 123 (2025) 102760

Contents lists available at ScienceDirect

Learning and Individual Differences

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



Multimethod assessment of self-regulated learning in primary, secondary, and tertiary education – A meta-analysis☆

Julia Ruhl *, Franziska Perels, Laura Dörrenbächer-Ulrich

Department of Educational Sciences, Saarland University, Saarbrücken, Germany

ARTICLE INFO

Keywords:
Self-regulated learning
Multimethod assessment
Meta-analysis

ABSTRACT

Self-regulated learning (SRL) can be measured in several ways, which can be broadly classified into online and offline instruments. Although both online and offline measurements have advantages and disadvantages, the over-dependence of SRL research on offline measurements has been criticised considerably. Currently, efforts are being made to use multimethod SRL assessments. We examined 20 articles with 351 effect sizes that assessed SRL with at least two instruments on at least two SRL components. Most effect sizes were not statistically significant but descriptively higher than others. Combinations of two online instruments showed the highest effect size (r = 0.24). Overall correlations between instruments were highest for university students (r = 0.21). Additionally, results for cognition showed the highest effect size measured with behavioural traces (r = 0.28), and for metacognition measured with microanalysis (r = 0.35). The component of motivation was best measured using self-report questionnaires (r = 0.29).

Educational relevance statement: Self-regulated learning is an important predictor of academical success. It is therefore necessary to measure it as precise and comprehensive as possible. Knowing which instruments are best suited for each age group, SRL component, or reliably predict a specific achievement variable can help educators pick the best instrument for their needs.

1. Introduction

Self-regulated learning (SRL) is important and necessary for academic success (Sitzmann & Ely, 2011) in primary (Throndsen, 2011), secondary (Benick et al., 2021), and tertiary (Kitsantas et al., 2008) education. Moreover, Wirth and Leutner (2008) call it a component of lifelong learning, and the European Framework of Life-long Learning (EU Council, 2002) states that the current society requires students to be able to learn in a self-regulated way during and after schooling and throughout their entire working life.

SRL as a construct emerged in the 1980s (Dinsmore et al., 2008), and with it, the need for its measurement. Early theories, which assumed SRL to be stable across contexts (Boekaerts & Corno, 2005), primarily focused on the metacognitive and cognitive components and developed questionnaires to measure it in as decontextualised manner as possible (Boekaerts & Corno, 2005; Entwistle, 1988; Flavell, 1979). However, current research is trying to move beyond this notion by including motivational components and using situation-specific assessment methods (Rovers et al., 2019). With self-report questionnaires becoming less favourable and new assessment methods emerging (Dinsmore et al., 2008), calls for multimethod assessment have been getting louder (Cleary & Russo, 2023; Cleary & Zimmerman, 2004). In recent years, multiple researchers have attempted to answer those calls – with varying methods and results. Some combined quantitative and qualitative measures (e.g., Jansen et al., 2020; Lau, 2012), whereas others used combinations of different quantitative instrument types simultaneously (e.g., DiBenedetto & Zimmerman, 2013; Follmer & Sperling, 2019). Some studies focused on SRL as a whole (e.g., Dörrenbächer-Ulrich et al., 2021), and others selected one component for analysis, such as motivation (Cleary & Kitsantas, 2017). The initial findings of comparing different instrument types indicate that there are relationships between the instruments as well as between the instruments and achievement (e.g., Dörrenbächer-Ulrich et al., 2021). However, some studies found those relations to be minimal or non-existent (e.g., Callan & Cleary, 2018).

Considering the importance of multimethod research for the future of SRL, we wanted to examine the studies that used such an approach. Therefore, the aim of this study is to determine if and how different

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

* Corresponding author at: Universität des Saarlandes, Campus A4.2, 66123 Saarbrücken, Germany.
E-mail address: julia.ruhl@uni-saarland.de (J. Ruhl).

https://doi.org/10.1016/j.lindif.2025.102760
Received 17 September 2024; Received in revised form 28 May 2025; Accepted 16 July 2025
Available online 6 August 2025

instruments relate to each other, whether some combinations produce stronger correlations than others, and how all of them relate to achievement as a whole and on component level.

1.1. Definition, development, and models of SRL

SRL is defined as a "process whereby learners activate and sustain cognitions, affects, and behaviours that are systematically oriented towards the attainment of personal goals" (Zimmerman & Schunk, 2011, p.1). Although different conceptualisations of SRL exist, it is generally agreed upon to consist of three different components: cognition, metacognition, and motivation (Perels et al., 2020; Boekaerts, 1999). The cognitive component includes knowledge about learning strategies, such as skills needed to encode, memorise, and recall information, as well as the ability to use those strategies (Perels et al., 2020; Shuy, 2010). Metacognition is necessary for understanding and monitoring one's learning process and includes planning, self-observation, reflection, and adaptive adjustments to one's learning behaviour (Perels et al., 2020; Veenman, 2013). The motivational component includes self-efficacy beliefs, self-motivation and volitional control, and beneficial causal attributions for success and failure (Perels et al., 2020). Over the years, different models have been developed to describe SRL as a concept and explain possible SRL processes (Panadero, 2017; Tinajero et al., 2024). Models by Boekaerts (1999, 2006, 2007) and Zimmerman (2000) are the most prominent and influential in SRL research with Zimmerman being one of the first SRL authors (Panadero, 2017; Zimmerman, 1986). Zimmerman adopted a socio-cognitive perspective and described SRL as a cyclical process consisting of three different but dependent phases (forethought, performance, and self-reflection), whereas Boekaerts' considered cognitive and motivational self-regulation to be the basic mechanisms of SRL (Panadero, 2017) and focused on distinguishing between the originally proposed components of cognition, metacognition, and motivation (Panadero, 2017; Tinajero et al., 2024). Later, models also included emotional and behavioural components as part of SRL (see Panadero, 2017).

When SRL first became a topic of interest for researchers, they focused on older student populations and assumed younger children to be incapable of any cognitive, metacognitive, or motivational regulation (Hutchinson et al., 2021). Currently, there is ample evidence that at least rudimentary SRL competencies exist in young school-aged children and even preschoolers (Jacob, 2020). During the preschool period, several developmental steps lead to the emergence of complex learning processes such as SRL (Jacob, 2020). Studies show that preschoolers can use different types of learning strategies, regulate their emotions, and enact metacognition (Hutchinson et al., 2021). A shift from emotion-driven regulation to more cognitive regulation has been observed during this stage (Zelazo, 2015). Nevertheless, SRL in young children remains understudied and usually focuses on confirming the existence of different phases according to Zimmerman (2000) and less on differentiating among the possible components of SRL (e.g., Heirweg et al., 2020). The first results suggest SRL to be only a one-dimensional structure in preschool children (Dörr & Perels, 2018) with multiple components; therefore, a multi-dimensional structure is first developed in primary-school age (Benick et al., 2018). By contrast, in older populations, such as high-school or university students, the existence of the three components has been studied and confirmed extensively and is the basis for several assessment methods (Dörrenbächer-Ulrich et al., 2024).

1.2. Assessment of SRL

Various instruments (see Table 1) have been developed to assess SRL as comprehensively as possible. Typically, they are categorised into two categories: those that consider SRL trait-like and measure it as learning behaviour on a more global level (*aptitude* or *offline measures*) and those that consider SRL to be a state and therefore assess it in a situation-specific manner (*event* or *online measures*; Winne, 2010; Wirth & Leutner, 2008). Past research considered SRL to be a stable individual characteristic, resulting in the dominant usage of trait-like measurements, such as questionnaires (Boekaerts & Corno, 2005; Endedijk et al., 2016). This reliance on offline measures has recently been challenged by scholars, leading to the development of more contextualised online measures that aim to measure SRL in real time (Boekaerts & Corno, 2005; Endedijk et al., 2016).

Table 1
Overview of different types of SRL instruments.

Instrument type	Name	Description
Offline/ aptitude measure	Self-report questionnaires	Learners rate statements about their abilities, behaviours, skills, etc.; retrospective and generalised
	Strategy knowledge tests	Learners rate the perceived usefulness of strategies in a previously presented situation independent of whether they would use the strategies themselves; the assumption that knowledge about strategies is a prerequisite for their usage
	Interviews	Similar to self-report questionnaires but with open-ended assessment instead of Likert-type rating
Online/event measure	Think-aloud protocol	Learners are asked to verbalise every thought while working on a task: no division into different phases of SRL
	Microanalysis	Learners work on a task divided into forethought, performance, and self-reflection; they are asked about their plans, behaviours, evaluations, etc. throughout the task
	Learning diaries	Learners are asked questions about their task-specific learning behaviour over a longer period of time
	E-portfolios	Similar to learning diaries but additional documentation such as assignments, self-reflections, etc. are collected
	Data traces, log files, eye tracking	Learners work on learning tasks while different types of information are collected, such as highlighting words, clicking on hyperlinks, or eye movements

Nevertheless, offline measures remain one of the most regularly used measurement types in SRL research although recently, their frequent usage has been criticised considerably (Dinsmore et al., 2008; Endedijk et al., 2016). They are generally considered to be poor indicators of students' actual SRL usage, and the focal point of criticism is that it remains unclear which point of reference students use to draw inferences (Dinsmore et al., 2008; Perry & Winne, 2006; Pintrich, 2004; van Hout-Wolters, 2000). Furthermore, there is concern about whether students interpret questionnaires as intended by the researcher or whether the surveys may induce responses that students would otherwise not report (Anthony et al., 2013; Brophy, 2005; Karabenick et al., 2007). Nevertheless, they work well for understanding students' general preferences regarding SRL and strategy usage (Endedijk et al., 2016). The most commonly used offline measures include self-report questionnaires (Roth et al., 2016; Schunk & Greene, 2018). In them, learners are presented with a variety of statements concerning aspects such as their general abilities, preferences, and behaviours and are asked to rate them on a Likert-type scale depending on how well the statements correspond to their experiences (Wolters & Won, 2018). Self-report questionnaires are very economical and can be easily administered to large groups of people (Rovers et al., 2019). Additionally, self-report questionnaires are standardised in implementation and interpretation, allowing for an objective assessment of SRL (Dörrenbächer-Ulrich et al., 2021). Their primary disadvantage is that they assess SRL retrospectively and in an aggregated manner over time, which can lead to both retention and generalisation problems (Rovers et al., 2019).

Furthermore, they are not very sensitive to changes in SRL behaviour and assess SRL as a trait-like characteristic, inadvertently assuming that it remains stable over time (Rovers et al., 2019). As mentioned previously, it remains unclear what type of situations students have in mind when responding to self-report questionnaires and whether they have any personal experiences in mind at all. Some authors suspect validity problems of self-report questionnaires as they combine the assessment of strategy knowledge and usage and may not reflect actual behaviour (Artelt, 2000; Dörrenbächer-Ulrich et al., 2021). Nevertheless, self-report questionnaires provide valuable information about learners' SRL, even if the results may not be completely accurate or correspond to behaviour (McCardle & Hadwin, 2015). Similar arguments can be applied to interviews, which are comparable to self-report questionnaires but typically consist of open-ended questions (Perels et al., 2020). A specific type of offline measure is the strategy knowledge test. Here, learners are presented with different possible SRL strategies for a specific type of situation and asked to rate the perceived usefulness of each strategy – independently of whether they themselves would use it (Perels et al., 2020). Test takers' consensus with expert ratings is evaluated as high strategy knowledge. This assessment type is deemed valid based on the assumption that knowledge about helpful SRL strategies is a prerequisite for their application (Dörrenbächer-Ulrich et al., 2024; Grassinger, 2011; Wolters, 2003). Similarly, to self-report questionnaires, strategy knowledge tests are economical and objective (Dörrenbächer-Ulrich et al., 2021). Thus far, they have been primarily used in and validated for German-speaking countries (Dörrenbächer-Ulrich et al., 2021).

Online measures are used less frequently in research (Dinsmore et al., 2008). Their primary and obvious advantage is their context-specificity and fine-grained measurement of SRL in real time (Endedijk et al., 2016). Their main disadvantage compared with offline measures is that they are unable to assess a learner's general SRL usage but can only consider behaviours shown in very specific types of situations; moreover, claims cannot be made about any other type of SRL strategy that a person may know or use (Endedijk et al., 2016). One task-specific way of assessing SRL is microanalytic assessment. Microanalysis is fine-grained and context-specific and measures SRL in real time during a specific task (Cleary, 2011). Microanalysis is a special form of think-aloud protocol (Perels et al., 2020), which asks a learner to verbalise every thought they have while working on a specific task. This allows for insight into the learner's spontaneous and specific strategy usage (Winne & Perry, 2000). Think-aloud protocols assess learning strategies during one continuous task and do not differentiate between different learning phases (Zimmerman, 2000). However, during microanalysis, learners work on a task that can be divided into the three phases postulated by Zimmerman (2000): forethought, performance, and self-reflection. Then, students are asked questions about planning before starting the task, questions on their current performance during the task, and to self-reflect after the task (Cleary & Callan, 2018). Microanalytic questions can be open- or close-ended. Their advantage over questionnaires is the task-specificity and, therefore, the lack of retrospective or prospective bias (Cleary et al., 2012). Research shows that microanalysis has good reliability and validity (Cleary et al., 2012) but low or no correlation with SRL questionnaires (Cleary et al., 2015; DiBenedetto & Zimmerman, 2013). This may be due to their differences in generalisation and specificity (Dörrenbächer-Ulrich et al., 2021). However, microanalysis shows associations with performance and achievement (Cleary et al., 2015; Lau et al., 2015). Another possible measure are learning diaries or e-portfolios (Perels et al., 2020; Dignath et al., 2023). Learning diaries ask the learner questions about a task-specific learning behaviour but over a longer period of time (e.g., daily over multiple weeks; Perels et al., 2020). *E*-portfolios additionally systematically collect students' work such as assignments, self-reflections, or other documents that contain information about the student's learning and progress (Chang et al., 2013). This allows to monitor if and when changes in learning behaviour occurred. Learning diaries/e-portfolios also foster learners' self-monitoring which is important for SRL (Dignath et al., 2023). A meta-analysis by Dignath et al. (2023) finds positive effects of these monitoring tools on SRL and academic achievement. The previously described online measures can be considered quite obtrusive to a person's learning as they require the learner's conscious attention (Siadaty et al., 2016). They demand the learner to stop what they are doing to answer questions about their learning process, which can disrupt a potential flow state of learning (Siadaty et al., 2016). As an alternative, more unobtrusive forms of online measures can be used that collect data in the background. Data traces, log files, or eye tracking are options for such measurements within online learning environments (Fan et al., 2022; Siadaty et al., 2016). All are based on the idea that two main types of temporal features are considered in learning analytics: how much time a person spends on a task, and in what order a person approaches different learning events (Fan et al., 2021). Trace data or log files can measure how much time a learner spends on a task, whether they highlight passages while reading, or how many provided hyperlinks they click on (Fan et al., 2021). All those actions provide information about a learner's cognitive and metacognitive activities during the learning task (Winne, 2011). Moreover, eye tracking can detect a learner's underlying cognitive and metacognitive monitoring during multimedia learning (Mayer, 2010; Mudrick et al., 2019; van Gog & Jarodzka, 2013). Eye movement can provide information about attentional processes and indicate meta-comprehension in case of discrepancies in a learning text (Jamet, 2014; Mudrick et al., 2019). Trace data especially has seen an emergence in popularity (Fan et al., 2022) as it promises a more authentic portrayal of a learner's SRL. This particular branch of SRL research sees its main challenge in the interpretation of the trace data as it is quite unclear how exactly a specific trace relates to a distinct SRL process (Du et al., 2023). Current guidelines suggest the interpretation of frequencies, transitions, and sequences, or combining trace date with other sources such as self-reports, think-aloud-data, or learning outcomes (Du et al., 2023; Fan et al., 2022; Siadaty et al., 2016).

Most instruments for measuring SRL require some degree of reading or verbalisation ability on the learner's part, making them suitable for learners starting in (the late stages of) primary school (Perels et al., 2020). This may be a reason many multimethod SRL studies focus on university students (e.g., Dörrenbächer-Ulrich et al., 2021), as with them, the options for SRL assessment are broader, and self-report questionnaires can be combined with more complex measures, such as traces (e.g., Hadwin et al., 2007; Zhou & Winne, 2012). Meanwhile, school-aged children are often evaluated using different types of (self-report) questionnaires (e.g., Benick et al., 2021; Chen et al., 2015; Cleary & Callan, 2013). Finally, observations can be used as an alternative assessment of preschool-aged children (Dörr & Perels, 2018; Perels et al., 2020). During observations, a trained observer rates the child's behaviours by answering a set of prepared questions (e.g., CHILD-Checklist; Anderson et al., 2003) with the option of taking additional notes. Observations are typically combined with video recordings so that multiple independent raters can score behaviours to increase the reliability of the instrument. As is typically the case with instruments that rely on third parties, the objectivity of observations is much lower than the objectivity of, for example, questionnaires (Perels et al., 2020). Newer methods of preschool SRL measurement are being developed, such as the dynamic assessment of self-regulated learning in preschool method (Moreira et al., 2022). The dynamic assessment of self-regulated learning in preschool considers all phases postulated by Zimmerman (2000) and combines interviews (during forethought and reflection) with observations of performance during a preschool-appropriate task (Zimmerman, 2000). So far, no research on the effect of age on the relationship between different instrument types exists. Taking the factorial differences in the structure of SRL (Dörrenbächer-Ulrich et al., 2024; Benick et al., 2018; Dörr & Perels, 2018) and differences in verbal abilities (and therefore the need for self-report alternatives; e.g., Perels et al., 2020; Benick et al., 2021; Chen et al., 2015; Cleary & Callan, 2013) into account, a moderating influence of age may be assumed

which could potentially manifest as higher correlations between instruments for older populations.

In general, neither type of instrument is superior to the other, with both offline and online measures showing clear advantages and disadvantages (Endedijk et al., 2016). The choice of measurement type depends on the context and nature of the research question, and all instruments can contribute to a better understanding of a person's SRL usage (Winne & Perry, 2000). However, efforts are being made to use a multimethod approach for SRL assessment, combining multiple online or offline measures or different types (e.g., Dörrenbächer-Ulrich et al., 2021; Callan & Cleary, 2018). Combining multiple instruments should increase the reliability of the assessment (Perry & Rahim, 2011; Veenman, 2011) and balance out the advantages and disadvantages of each assessment method.

1.3. Multimethod assessment: relationships between different instrument types and relations to achievement

Although the theoretical arguments for multimethod assessment seem sound, the empirical results are not as straightforward. Various studies have examined multimethod SRL assessment from different perspectives (i.e., by using different combinations), and the results vary (e.g., Callan & Cleary, 2018; Chen et al., 2015; Cleary et al., 2015). Combinations of the same measurement categories (online/online, offline/offline) tend to produce statistically significant correlations, whereas combinations of different categories (online/offline) tend to perform poorly (Callan & Cleary, 2018). Self-report questionnaires and teacher/parent ratings have strong relationships with each other, as do two event measures, such as microanalysis and behavioural traces (Callan & Cleary, 2018; Chen et al., 2015). Similar has been found for questionnaires and interviews (Anthony et al., 2013). However, many studies don't find correlations between questionnaires and any event measures (Cleary et al., 2012, 2015; Veenman et al., 2003; Winne & Jamieson-Noel, 2002), or between teacher ratings and event measures (DiBenedetto & Zimmerman, 2013). Nevertheless, some exceptions have been noted. Callan and Cleary (2018) found correlations between teacher ratings and microanalytic measures of metacognitive monitoring, and Dörrenbächer-Ulrich et al. (2021) found correlations between questionnaires and microanalysis for the motivational component. Regarding trace data, van Halem et al. (2020) found that strategy use scales of the MSLQ significantly predict study behaviour (measured with behavioural traces). Additionally, specific and global measures of strategy use do not correspond to each other (Callan & Cleary, 2018). This may be because different contexts require different strategies, and therefore, strategy use tends to vary across different circumstances (Lodewyk et al., 2009). Overall, the poor correlative results between instrument types may be because different SRL measures target different SRL aspects (Callan & Cleary, 2018). Given that some studies only found correlative results for some components (e.g., Anthony et al., 2013; Dörrenbächer-Ulrich et al., 2021; Callan & Cleary, 2018), it can be hypothesised that different components of SRL are influenced by different characteristics, which, in turn, are best measured using different instruments (Dörrenbächer-Ulrich et al., 2021). Cognition is highly influenced by the learner's knowledge of learning strategies (Boekaerts, 1999) and may, therefore, be best measured using strategy knowledge tests (e.g., Dörrenbächer-Ulrich et al., 2021). Likewise, the application of those learning strategies is equally important (Zimmerman, 2008) and may in turn be best measured using trace or think-aloud data (e.g., Azevedo et al., 2010). Motivation consists of different beliefs (Pintrich, 2004), and beliefs are generally best assessed using questionnaires (Pintrich et al., 1993). Further, long-term motivation is considered to be most beneficial for SRL (Zimmerman, 2000) which may also speak in favour for an assessment using questionnaires or instruments like learning diaries (e.g., Zimmerman, 2008) which can take multiple learning situations over a longer period of time into account. Motivation could also be very well observed. Observational measures, such as the CHILD-checklist (Anderson et al., 2003) take different forms of observable behaviour into account which might speak for high levels of motivation, such as initiating activities or enjoying problem solving (Anderson et al., 2003; Perels et al., 2020). Finally, metacognition is the most situation-specific of all components and, therefore, may be best assessed using event measures, such as microanalytic assessment (e.g., Dörrenbächer-Ulrich et al., 2021; Callan & Cleary, 2018, DiBenedetto & Zimmerman, 2013). It can also be assessed, and is frequently being assessed, using self-report questionnaires such as the MSLQ (e.g., Gandomkar et al., 2020; Pintrich et al., 1993; Sperling et al., 2004).

Moreover, relations to or predictions of achievement vary depending on the SRL instrument used. Although both online and offline measures have been shown to predict achievement, the findings are mixed (Callan & Cleary, 2018). Some studies found event measures to be better predictors of achievement variables such as grades (Cleary et al., 2015; DiBenedetto & Zimmerman, 2013), whereas others show aptitude measures to be more robust (Cleary & Chen, 2009; Jamieson-Noel & Winne, 2003). Other studies found both assessment types to relate to achievement with no clear indication of one being superior (Dörrenbächer-Ulrich et al., 2021; Gandomkar et al., 2020). Furthermore, the predictive validity of the instruments may depend on the type of achievement variable measured. Callan and Cleary (2018) assumed event measures to be better predictors for more contextualised outcomes and aptitude for more general ones, meaning that event or online measures correlate higher with performance measured during a study (e.g., a multiple-choice test that was developed to be used during microanalysis) while aptitude or offline measures correlate higher with performance shown outside the context of a study (e.g., school grades). Their data support this theory: microanalysis metacognitive monitoring was the strongest predictor of the performance in practice session math problems, whereas teacher ratings best predicted the more global standardised math test achievement measure (Callan & Cleary, 2018). Other studies support those findings (Callan & Cleary, 2019; Veenman & Van Cleef, 2019). However, some exceptions have been found using microanalytic metacognitive monitoring (Callan & Cleary, 2018) as well as observational and think-aloud data (Veenman & Van Cleef, 2019), relating to both specific and global outcomes (Gandomkar et al., 2020).

1.4. Aim of the present study

To summarise, there is a call (e.g., Dörrenbächer-Ulrich et al., 2021; Callan & Cleary, 2018) for multimethod assessment in the current SRL research. Initial research has shown poor correlative results between some SRL instruments, with few exceptions at the level of different SRL components (Dörrenbächer-Ulrich et al., 2021; Callan & Cleary, 2018). Therefore, it is theorised that some measures may target certain SRL components better than others (Dörrenbächer-Ulrich et al., 2021). For example, metacognition may be best assessed using microanalysis, as it is the most situation-specific component among the three; motivation by using self-report questionnaires, as they can consider multiple situations; and cognition with strategy knowledge tests, as cognition primarily benefits from knowing different learning strategies (Dörrenbächer-Ulrich et al., 2021). Additionally, differences between SRL instruments regarding their relations to achievement have been observed, with the most notable theory suggesting that event measures are better predictors for more contextualised outcomes and aptitude for more general ones (Callan & Cleary, 2018). Furthermore, although the developmental characteristics of SRL and the usage of different instruments in different age groups suggest an influence of age on instruments, no research thus far has focused on a potential moderating effect of age on the correlation between SRL instruments.

Therefore, the current meta-analysis has the following goals. First, we examine studies that have used a multimethod approach to assess SRL and analyse how the different instruments relate to each other and achievement. Then, while considering achievement variables, we

analyse whether certain instruments are better suited to assess a specific SRL component. For this, we explore correlations between instrument types and achievement for each component individually. Finally, we test whether age influences the correlation between instruments. This leads to the following research questions and hypotheses:

1. How do different types of SRL instruments/measurements relate to each other?
   → H1: The correlations of the same types of instruments (online/online or offline/offline) are higher than those of combinations of different types of instruments.
2. Do some instruments correlate higher with achievement?
   → H2: Online instruments correlate higher with study-specific achievement (e.g., task performance), and offline measures correlate higher with general academic achievement (e.g., grades).
3. Are certain instruments better suited to assess a specific SRL component (cognition, metacognition, motivation) in terms of their correlation to achievement measures as a validity criterion?
   → H3: Cognition is more validly assessed using strategy knowledge tests, metacognition using microanalysis, and motivation using self-report questionnaires.
4. Is age a moderator of the correlation between different SRL instruments? (exploratory)

2. Method

2.1. Search strategy

All databases available on EBSCOhost were searched in spring 2024. Eleven separate searches had to be performed due to the complexity and number of search terms. The following search terms were always included: "*self-regulated learning*" OR *SRL* AND *achievement* OR *performance* OR *success* AND *assessment* OR *measur** NOT *preschool*. We decided to exclude papers on preschools from our analyses as they usually require very different assessment types than other populations (Perels et al., 2020); more importantly, research thus far suggests that the SRL of preschoolers is unidimensional and not divisible into the three components (Dörr & Perels, 2018). Individual search terms for searches n1 to n11 can be seen in Table 2. The search was limited to include articles published between 2003 and 2024. The goal was to cover the past 20 years of research. SRL models by Boekaerts and Zimmerman were introduced in 1999 and 2000, respectively. As they form the basis of our SRL definition, we wanted to include all possible articles from after the models' introduction. We assumed a short gap between the introduction of the models and actual research being published that focuses on them. Early studies on SRL also mostly used self-report questionnaires with variance in measurements only getting introduced later on (Dinsmore et al., 2008). We therefore chose 2003 as a starting point. The literature search was conducted in 2023 and repeated in 2024.

Overall, 7228 articles were found. One hundred and twenty-eight articles were removed because they were written in a language other than English or German. The remaining articles were equally divided between two researchers and screened on the title and abstract level. During this step, we discarded everything that was evidently irrelevant to our research questions, such as papers from unrelated non-pedagogical/non-psychological fields. This led to another 7003 articles being removed, mostly due to being from unrelated fields, such as chemistry and physics, and being irrelevant to SRL. Finally, 97 articles remained and were sought for retrieval. This revealed 32 duplicates, which were removed. The remaining articles were screened on a full-text level by two trained researchers. Interrater reliability was $\kappa = 0.83$. In case of discrepancies, the decision for inclusion or exclusion lay with the first author. Ultimately, the following inclusion criteria had to be met for articles to be included in the analysis: First, they had to be empirical studies that measured SRL as a whole on at least two components (cognition, metacognition, motivation). We decided to focus on these three components, as they are most frequently included in definitions of SRL and were also postulated by Boekaerts (1999) in her model. Some more recent reviews suggest that SRL additionally also consists of emotional and behavioural components (Panadero, 2017; Zeidner & Stoeger, 2019). As those are not part of the original three and get mentioned first in later articles, we decided to not specifically make them part of our inclusion criteria. Furthermore, they had to focus on student samples from primary, secondary, or tertiary education while

Table 2
Search terms for searches n1–n11.

Topic (number of results)	Label	Search term
		("self-regulated learning" OR SRL) AND (achievement OR performance OR success) AND (assessment OR measur*) NOT (preschool)
Questionnaire ($n = 4147$)	n1	((questionnaire AND interview) OR (questionnaire AND "strategy knowledge test") OR (questionnaire AND "teacher ratings") OR (questionnaire AND "trace data") OR (questionnaire AND microanalysis) OR (questionnaire AND "log files") OR (questionnaire AND observation) OR (questionnaire AND "eye tracking") OR (questionnaire AND "thinking aloud") OR (questionnaire AND "learning diary") OR (questionnaire AND eportfolio))
Interview ($n = 2319$)	n2	((interview AND "strategy knowledge test") OR (interview AND "teacher ratings") OR (interview AND "trace data") OR (interview AND microanalysis) OR (interview AND "log files") OR (interview AND observation) OR (interview AND "eye tracking") OR (interview AND "thinking aloud") OR (interview AND "learning diary") OR (interview AND eportfolio))
Strategy knowledge test ($n = 9$)	n3	(("strategy knowledge test" AND "teacher ratings") OR ("strategy knowledge test" AND "trace data") OR ("strategy knowledge test" AND microanalysis) OR ("strategy knowledge test" AND "log files") OR ("strategy knowledge test" AND observation) OR ("strategy knowledge test" AND "eye tracking") OR ("strategy knowledge test" AND "thinking aloud") OR ("strategy knowledge test" AND "learning diary") OR ("strategy knowledge test" AND eportfolio))
Teacher ratings ($n = 98$)	n4	(("teacher ratings" AND "trace data") OR ("teacher ratings" AND microanalysis) OR ("teacher ratings" AND "log files") OR ("teacher ratings" AND observation) OR ("teacher ratings" AND "eye tracking") OR ("teacher ratings" AND "thinking aloud") OR ("teacher ratings" AND "learning diary") OR ("teacher ratings" AND eportfolio))
Trace data ($n = 94$)	n5	(("trace data" AND microanalysis) OR ("trace data" AND "log files") OR ("trace data" AND observation) OR ("trace data" AND "eye tracking") OR ("trace data" AND "thinking aloud") OR ("trace data" AND "learning diary") OR ("trace data" AND eportfolio))
Microanalysis ($n = 104$)	n6	((microanalysis AND "log files") OR (microanalysis AND observation) OR (microanalysis AND "eye tracking") OR (microanalysis AND "thinking aloud") OR (microanalysis AND "learning diary") OR (microanalysis AND eportfolio))
Log files ($n = 77$)	n7	(("log files" AND observation) OR ("log files" AND "eye tracking") OR ("log files" AND "thinking aloud") OR ("log files" AND "learning diary") OR ("log files" AND eportfolio))
Observation ($n = 336$)	n8	((observation AND "eye tracking") OR (observation AND "thinking aloud") OR (observation AND "learning diary") OR (observation AND eportfolio))
Eye tracking ($n = 40$)	n9	(("eye tracking" AND "thinking aloud") OR ("eye tracking" AND "learning diary") OR ("eye tracking" AND eportfolio))
Thinking aloud ($n = 2$)	n10	(("thinking aloud" AND "learning diary") OR ("thinking aloud" AND eportfolio))
Learning diary ($n = 2$)	n11	(("learning diary" AND eportfolio))

not centring gifted students or students with special needs. SRL had to be measured using at least two different instruments, and some type of achievement/performance variable, such as grades, had to be included. This led to 22 eligible articles. Another nine eligible articles were found by looking through works from known SRL researchers and by reading through two other meta-analyses on the topic of SRL interventions (Theobald, 2021; Theobald et al., 2023). Finally, only 20[1] articles (18 from the original search and two from the additional search) could be included in the analysis as the other articles didn't report the necessary statistics and messaging the authors didn't yield any results. An overview of the articles can be found in Table 3. The number of articles included and excluded from every search category in every step of the process can be seen in Fig. 1.

2.2. Data extraction and preparation

A coding sheet was developed containing the following information: study ID, publication year, country, sample size, number of female participants, current level of education, mean age and standard deviation age. Moreover, we included type of SRL instruments used in the study and their reliability, mean, and standard deviation; SRL values for each component; type of achievement/performance measure(s) and their mean and standard deviation; correlations between SRL instruments as well as correlations between SRL and performance. The development of the coding sheet was discussed with the second author. Data extraction was performed by the first author. First, all basic information, such as publication year, country, sample size, and the participants' mean age, was extracted. Furthermore, all instruments used to assess SRL and ways to assess achievement/performance used in the studies were extracted. Finally, relevant correlation and regression coefficients (such as correlations between the SRL instruments or between SRL and achievement) as well as effect sizes (e.g., Cohen's *d* or similar) were extracted. Problems and uncertainties during data extraction were discussed with the second author.

All effect sizes were converted into Pearson's *r*, as this was the most commonly used effect size across studies. Two studies reported Kappa (κ) and η^2, respectively. According to literature, κ equals Pearson's *r* (Rettew et al., 2009) and η^2 equals r^2 (Lakens, 2013). Further, all SRL instruments were categorised as either online or offline instruments. The instruments assessing achievement were categorised as either general academic (e.g., grades) or study-specific (e.g., study performance) assessment methods. Relevant analyses are reported separately for each research question.

2.3. Analyses

Data analysis was performed using the open-source software R with the package *robumeta* (Fisher & Tipton, 2015) which is a package for conducting robust variance estimations (RVE) in small and large samples. Data was analysed using RVE due to dependency in the data caused by multiple effect sizes being included from most studies (Fisher & Tipton, 2015; Hedges et al., 2010). RVE are a way to adjust for heterogeneity and issues in the estimation of standard errors when the assumption of independent observations is violated. We assumed a correlated effects model, to account for within-study correlations, and used small sample adjustments (Tipton, 2015). Generally, RVE are designed to be used in meta-analyses with large samples and are able provide an unbiased estimator of the true sampling variance (Tipton & Pustejovsky, 2015). In smaller samples, the Type I error rate of hypothesis tests based on uncorrected RVE tends to be too liberal (Hedges et al., 2010; Tipton & Pustejovsky, 2015). To counteract, Tipton (2015) proposes small sample corrections which can be implemented in meta-analyses with as few as five studies and is recommended for meta-analyses containing fewer than 50 studies (Tipton, 2015; Tipton & Pustejovsky, 2015). Additionally, moderator analyses were calculated using the same package. All effect sizes *r* were converted into Fisher's Z (Borenstein et al., 2009), and variances were calculated based on the standardised values as well. For the presentation of the results, all values were converted back into *r*. Forest plots were made using the function forest.robu (Fisher & Tipton, 2015) and generated based on the original correlational coefficients. For the examination of study influence, small study bias, and publication bias we used the *metafor* package (Viechtbauer, 2021). Analyses will be described in more detail for each research question.

3. Results

3.1. Publication bias, small study bias, study influence and power-analysis

Publication bias was assessed using the *metafor* package (Viechtbauer, 2021). A multilevel Egger-like regression test was conducted to examine potential publication bias, accounting for the hierarchical structure of the data (i.e., multiple outcomes nested within studies). The model included the standard error as a moderator and random effects at the study level and the nested outcome level. The results revealed a significant negative association between effect size and its standard error, $\beta = -5.78$, $SE = 0.35$, $z = -16.64$, $p < .001$, 95 % CI [−6.46, −5.10], indicating evidence of small-study effects. The test of moderators was significant, $Q(1) = 277.05$, $p < .001$. Residual heterogeneity was substantial, $Q(349) = 2407.49$, $p < .001$. These findings suggest the presence of small-study effects, which may indicate publication bias or other related biases. However, it has been suggested that the Egger's test does not provide valid results in the case of dependent effect sizes (Park et al., 2025). Furthermore, in the case of dependent effect sizes, publication bias cannot be handled using trim-and-fill procedures as this method does not produce reliable results (Peters et al., 2007; Terrin et al., 2003). We, therefore, re-ran analyses using PET regression with RVE. The PET-PEESE method (Stanley & Doucouliagos, 2014) did not find evidence of small study bias ($p = .540$). The difference in results might be because of the violation of independent effect sizes when using the Egger's test (Park et al., 2025), but it might also be that the PET-PEESE performed badly due to the small sample size (20 studies is defined as the cut-off; Stanley, 2017). Ultimately, due to both the presence of dependent effect sizes and the fact that the sample size is very small, there is no way to test and correct for publication bias that will produce fully valid results (Harrer et al., 2021). Therefore, the following results have to be interpreted with the knowledge of a possible small study bias.

Using the results of the Egger's test we also tested our data for outliers for each research question separately. There were no outliers (defined as data twice larger than the mean effect size) in our data.

A post-hoc power-analysis conducted with the *metapower* package revealed, that assuming an effect size of $r = 0.3$ with a heterogeneity of $I^2 = 75$ % in an average sample of $n = 50$, ten studies would be enough in a random-effects model to find an effect with a power of 0.91. For subgroup analysis with 3 subgroups and assuming effect sizes of $r = 0.1$, $r = 0.3$, and $r = 0.4$, depending on the subgroup, and an average sample size of $n = 60$ per study, with 20 studies one would find an effect with a power of 0.47.

3.2. Research question 1: How do different types of SRL instruments/measurements relate to each other?

Data were prepared by extracting correlations between different types of instruments. For every correlation, the instrument comparison

[1] Four of them did not fulfil the pre-registered requirement of including an achievement variable but were included to increase the sample size for research questions 1 and 4, as information about achievement is not necessary to analyse those research questions.

Table 3
Overview of analysed articles.

Author (year)	Nation	Sample size	Age/grade participants	Research design	SRL instruments	Performance variable+	SRL component/ outcome*
Anthony et al. (2013)	USA	160	9th (n = 81) and 10th (n = 74) grade girls	Mixed-methods	Microanalysis, self-report questionnaires	English final grade, math final grade	All SRL components with English and math grades
Bellhäuser et al. (2023)	Germany	194	College students, Mean age M = 22.21 (male n = 96, female n = 117)	Experimental	Self-report questionnaire, learning diary	n.a.	n.a.
Callan and Cleary (2018)	USA	100	8th grade	Correlational	Microanalysis, behavioural traces, self-report questionnaires, teacher ratings	Mathematical problem-solving skill, a global measure of mathematical skill	Cognition and metacognition with mathematical problem-solving skill and global measure of mathematical skill
Callan and Cleary (2019)	USA	96	8th grade students (male n = 42, female n = 54)	Correlational	Microanalysis, behavioural traces, interviews	Prior math achievement, practice session achievement	All components with practice session achievement
Chen et al. (2015)	USA	445	241 of 6th graders (male n = 119, female n = 122) and 204 of 7th graders (male n = 102, female n = 102)	Correlational	Self-report questionnaires, parent ratings	Math grades	Cognition and metacognition with math grades
Cleary and Callan (2013)	USA	87	9th grade students (56 % female)	Correlational	Self-report questionnaires, teacher ratings	Measure of Academic Progress (MAP; prior mathematic skill), math classroom test percentage	All components with prior mathematic skills and math classroom test percentage
Cleary et al. (2015)	USA	363	6th and 7th grade students (56 % female)	Correlational	Self-report questionnaires, teacher ratings	n.a.	n.a.
Dörrenbächer and Perels (2016)	Germany	174	College students, Mean age M = 22.94 (72 % female), n = 44 in relevant group	Quasi-experimental	Self-report questionnaire, learning diary	GPA (reverse coded), working efficiency test	All components with GPA and working efficiency test
Dörrenbächer-Ulrich et al. (2021)	Germany	70	College students, Mean age M = 22 (77.50 % female)	Correlational	Microanalysis, trace data, strategy knowledge test, SRL questionnaire	GPA (reverse coded)	All components with GPA
Dörrenbächer-Ulrich et al. (2024)	Germany	143 99 T1: 207 T2: 105	Pilot Study: Teacher education students, mean age M = 24.8 (70.3 % female) Validation Study I: Teacher education and psychology students, mean age M = 21.33 (75.8 % female) Validation Study II: Teacher education students, mean age t1: M = 21.86, t2: M = 21.69 (t1 = 76.8 % female, t2 = 80.00 % female)	Correlational	Strategy knowledge test, self-regulated learning questionnaire, microanalytic assessment	GPA (reverse coded)	All components with GPA
DiBenedetto and Zimmerman (2013)	USA	51	11th grade students (male n = 17, female n = 34)	Correlational	Microanalysis, interviews	Prior science achievement, tornado knowledge test, conceptual model test	Cognition and metacognition with prior science achievement, tornado knowledge test, and conceptual model test
Fan et al. (2023)	Netherlands	44	College students, mean age M = 21.70	Correlational	Trace data, think-aloud-data	n.a.	n.a.
Follmer and Sperling (2019)	USA	32	Undergraduate educational psychology course, mean age M = 19.68 (62.50 % female)	Correlational	Microanalysis, self-report questionnaires, strategy knowledge test	GPA, reading comprehension	(Cognition and) metacognition with reading comprehension
Gandomkar et al. (2020)	Iran	76	First-year medical students	Correlational	Microanalysis, behavioural traces, interviews, trace data, strategy knowledge tests, self-report questionnaires	Biomedical science course performance, learning task performance	All components with learning task performance and course performance
Maag Merki et al. (2013)	Switzerland	2300	Secondary school students	Correlational	Strategy knowledge test, self-report questionnaires	Self-reported German grades	All components with German grades
Metallidou and Vlachou (2010)	Greece	263	5th and 6th grade students (n = 133 female)	Correlational	Self-report questionnaires, teacher ratings	n.a.	n.a.

(*continued on next page*)

Table 3 (*continued*)

Author (year)	Nation	Sample size	Age/grade participants	Research design	SRL instruments	Performance variable[+]	SRL component/ outcome*
Paans et al. (2018)	Netherlands	62	5th grade students ($n = 38$ female), mean age $M = 10.00$), only $n = 16$ analysed for both log files and think-aloud	Correlational	Log files, think-aloud protocol	Knowledge test	Cognition and metacognition with knowledge test
Torrington et al. (2023)	Australia	48	Elementary school students	Correlational	Think-aloud protocol, self-report questionnaire	Written academic output (research task)	All components with task performance
van Halem et al. (2020)	Netherlands	605	First-year students: 2016: $n = 435$, mean age $M = 20.6$, 94.44 % female 2017: $n = 489$, mean age $M = 20.88$, 78.90 % female	Correlational	Self-report questionnaire, online trace data, multiple-choice-questionnaire	Grade multiple choice exam, grade research report	All components with grade multiple choice exam
Zhidkikh et al. (2023)	Finland	20	8th grade students	Correlational	Self-report questionnaires, log files, observations	Graded assignments	All components with graded assignments average

Note: *which SRL components were analysed regarding which outcomes. n.a. = not applicable. [+]For studies by German researchers investigating relationships with grades (which are reverse coded in Germany) we used the absolute value in analyses but the original negative correlations in the forest plots.

was assigned. Thus, if the correlation was between two online instruments, it was labelled as '1', between two offline instruments as '2', and between one online and one offline instrument as '3'. Coefficients that were not Pearson's *r* were converted. For most studies, this procedure produced multiple effect sizes. Multiple effect sizes from each study mean that correlation independence could not be assumed. Dependencies in effect sizes can reduce heterogeneity, known as the unit-of-analysis error (Harrer et al., 2021). To counteract this, robust variance estimations were calculated (Hedges et al., 2010). We assumed a correlated effects model and made small sample adjustments as we had less than 50 studies (Tipton, 2015). We analysed all 20 available articles, some of which were divided into multiple studies, which means that the analyses were run with $n = 28$ studies and $k = 351$ effect sizes. The robust variance estimations *t*-test was significant ($t(26) = 6.55$, $p < .01$) which suggests that there is a relationship between the instrument types.

To check whether there are differences between the instrument groups, moderator analyses were carried out using the *robumeta* package. The results for each moderator level are presented in Table 4. We used the *clubSandwich* package (Pustejovsky et al., 2025) to conduct pairwise comparisons to test whether there were any statistically significant differences between the three moderator levels.

Overall, results showed small to medium effect sizes (Cohen, 1992) for any type of instrument combination. In our sample of studies, a combination of two online instruments was descriptively larger with $r = 0.24$. The differences between the subgroups were not statistically significant. Effect sizes for offline/offline ($r = 0.14$, $t(3.53) = 3.75$, $p < .01$) and online/offline ($r = 0.15$, $t(4.61) = 3.46$, $p < .01$) are significantly different from zero suggesting that they each explain some variance across studies. The reason online/online isn't significantly different from zero might be due to large variability in the effect size.

3.3. Research question 2: Do some instruments correlate higher with achievement?

Data were prepared by extracting Pearson's *r* correlations between the SRL instrument and the measure of achievement/performance. For one study, Cohen's *d* had to be converted using the formula postulated by Ruscio (2008). One article had to be excluded as it was impossible to extract the necessary correlations. Another four articles were excluded as they did not report any achievement variables (see Footnote 2). Furthermore, the data were coded depending on whether an online or offline measure was used, and whether it was examined in combination with general (academic) or context-specific (study) performance. This led to the emergence of the following categories: 1 = online + study, 2 = online + academic, 3 = offline + study, and 4 = offline + academic. Some articles were divided into multiple studies as they looked at multiple categories. Given that again, we had multiple effect sizes for each study and, therefore, effect size independence could not be assumed, we conducted robust variance estimations. The forest plot is presented in Fig. 2.

Overall, the analysis of research question 2 included 15 articles with $n = 31$ studies, resulting from dividing some articles into multiple studies, and $k = 195$ effect sizes. The robust variance estimations *t*-test was significant ($t(28.9) = 5.68$, $p < .001$), suggesting a significant relationship in the data. Overall effect size was $r = 0.25$ ($SE = 0.04$, 95 % CI [0.16, 0.33]). Statistics for heterogeneity were $I^2 = 87.60$ and $\tau^2 = 0.05$, which is quite high.

To check whether there are differences between the instrument groups, moderator analyses were carried out. Results for each subgroup are listed in Table 5. Pairwise comparisons were calculated to analyse whether the subgroups were statistically different from each other.

Overall, the results showed medium effect sizes for online and offline measures combined with general academic achievement as well as offline measures combined with study-specific achievement (Cohen, 1992). In our sample of studies, measuring general academic achievement with online measures produced the descriptively highest correlation ($r = 0.25$) followed by study-specific achievement measured with offline measures ($r = 0.24$). The between subgroup differences were not statistically significant. The effect sizes for online + academic ($r = 0.25$, $t(5.29) = 3.99$, $p < .01$), offline + study ($r = 0.24$, $t(1.59) = 8.87$, $p < .05$), and offline + academic ($r = 0.20$, $t(5.70) = 4.31$, $p < .01$) were significantly different from zero. This means that these moderators each individually explain some of the variance across studies but cannot be statistically distinguished from each other.

3.4. Research question 3: Are certain instruments better suited to assess a specific SRL component (cognition, metacognition, motivation) in terms of their correlation to achievement measures as a validity criterion?

This research question was answered by conducting three smaller meta-analyses. This approach may not fully answer the research question, as it is complex and difficult to operationalize. It should, therefore, only be seen as a first attempt to approach the answer to this question.

Data were sorted depending on whether cognition, metacognition, or motivation were measured. To operationalize which instrument is "better", we looked at the correlation between the instrument types and performance. There was no differentiation between general or specific performance. Four studies had to be excluded due to not reporting any performance/achievement variable (see Footnote 2). Data were coded depending on the instrument type with 1 = self-report, 2 = third-party rating, 3 = strategy knowledge test, 4 = microanalysis, 5 = behavioural traces, 6 = observation, 7 = learning diary, and 8 = think-aloud. This

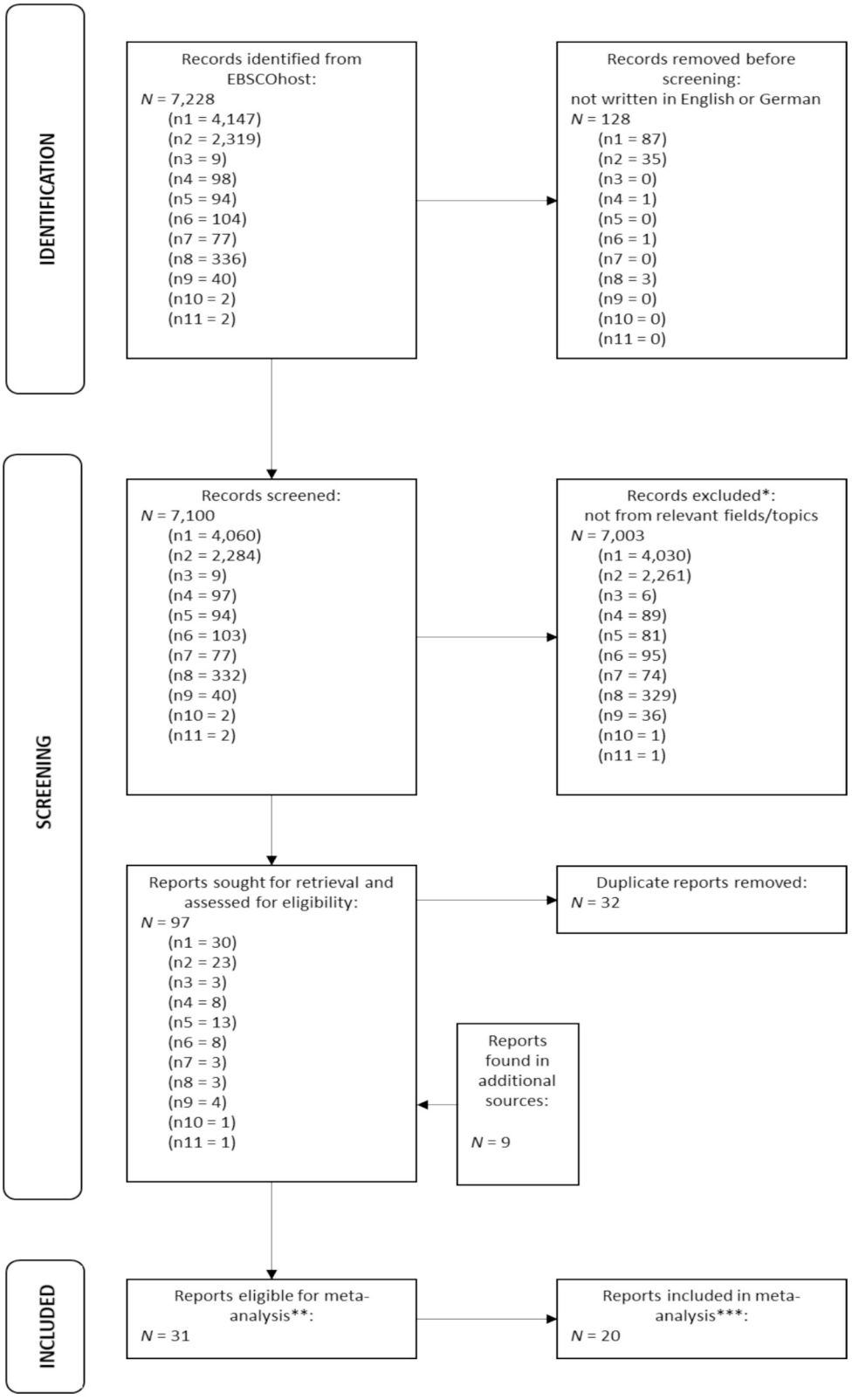


Fig. 1. Records identified and excluded at each screening stage.
Note. * based on inclusion and exclusion criteria on the title or abstract level.
** based on inclusion and exclusion criteria on the full-text level.
*** reports were excluded if they didn't report the relevant statistics and data could not be obtained by other means (e.g., open-source data, messaging authors).
n1–n11 refer to the topic-related search terms (Table 2).

Table 4
Moderator analysis for research question 1.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Online/online	5	77	0.24	[−0.15, 0.57]	0.02	0.15
Offline/offline	12	72	0.14	[0.03, 0.25]	0.02	0.15
Online/offline	11	202	0.15	[0.03, 0.26]	0.02	0.15

Note. n = number or studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

also led to some articles, that took different instruments into account, being divided into multiple studies. Sometimes, instruments were represented by only a handful of articles, or in extreme cases by only one. We decided to only include instruments (and therefore effect sizes) if there were at least three articles looking at that specific instrument/component combination. This led to excluding 1–3 more articles, depending on the component. It is also to note, that not every single article looked at all three components as some only considered two in their analyses. This also contributed to a significantly lower sample size depending on the component. As most studies reported more than one effect size, we once again calculated robust variance estimations.

For the component *cognition*, this led to the inclusion of 12 articles with $n = 21$ studies and $k = 40$ effect sizes. The forest plot is presented in Fig. 3. The robust variance estimations *t*-test was significant ($t(18.2) = 4.28$, $p < .001$). Overall effect size was $r = 0.17$ ($SE = 0.04$, 95 % CI [0.09, 0.25]). Statistics for heterogeneity were $I^2 = 79.99$ and $\tau^2 = 0.01$.

Moderator analyses were carried out for the following instrument subgroups: self-reports, strategy knowledge tests, microanalysis, and behavioural traces. Results for each subgroup can be seen in Table 6. Pairwise comparisons were calculated to determine whether the subgroups differed statistically from one another.

Overall, results showed small to medium effect sizes (Cohen, 1992) for strategy knowledge tests, microanalysis and behavioural traces and correlations close to zero for self-report measures. In our sample of studies, behavioural traces produce the highest effect size for the component of cognition with $r = 0.28$, which is also significantly different from zero ($t(3.19) = 8.81$, $p < .05$). None of the pairwise comparisons are statistically significant. This means that behavioural traces significantly predict the outcome variable but cannot be statistically distinguished from the other moderators. This might be due to insufficient power or collinearity/overlaps between the moderators (Lipsey, 2003).

For the component *metacognition*, this led to the inclusion of 13 articles with $n = 20$ studies and $k = 51$ effect sizes. The forest plot is presented in Fig. 4. The robust variance estimations *t*-test was significant ($t(16.4) = 4.61$, $p < .001$). Overall effect size was $r = 0.16$ ($SE = 0.04$, 95 % CI [0.09, 0.24]). Statistics for heterogeneity were $I^2 = 74.08$ and $\tau^2 = 0.01$.

Moderator analyses were carried out for the following instrument subgroups: self-reports, strategy knowledge tests, and microanalysis. Results for each subgroup can be taken from Table 7. Pairwise comparisons were calculated to analyse whether the subgroups were statistically different from each other.

Overall, results showed small effect sizes (Cohen, 1992) for self-report questionnaires ($r = 18$) and strategy knowledge tests ($r = 0.16$) and a descriptively medium effect size for microanalysis with $r = 0.35$. None of the pairwise comparisons were statistically significant. The

Table 5
Subgroup analysis for research question 2.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Online + study	8	81	0.04	[−0.34, 0.41]	0.05	0.23
Online + academic	8	33	0.25	[0.09, 0.39]	0.05	0.23
Offline + study	4	31	0.24	[0.12, 0.35]	0.05	0.23
Offline + academic	11	50	0.20	[0.09, 0.31]	0.05	0.23

Note. n = number of studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

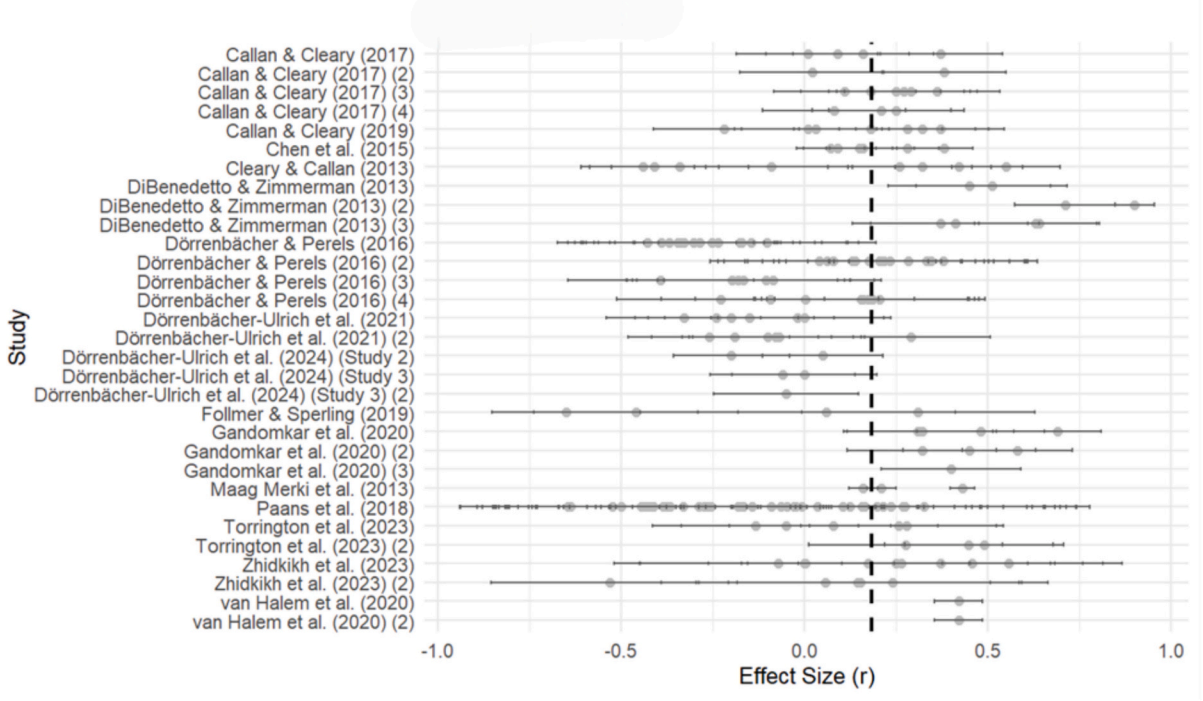


Fig. 2. Forest plot for research question 2.
Note. Forest plots were created using the original effect sizes from the studies while effect sizes for RVE and moderator analyses were Fisher *Z*-transformed beforehand. For studies by German researchers investigating relationships with grades (which are reverse coded in Germany) we used the absolute value in analyses but the original negative correlations in the forest plots (see Table 3 for which studies are affected).

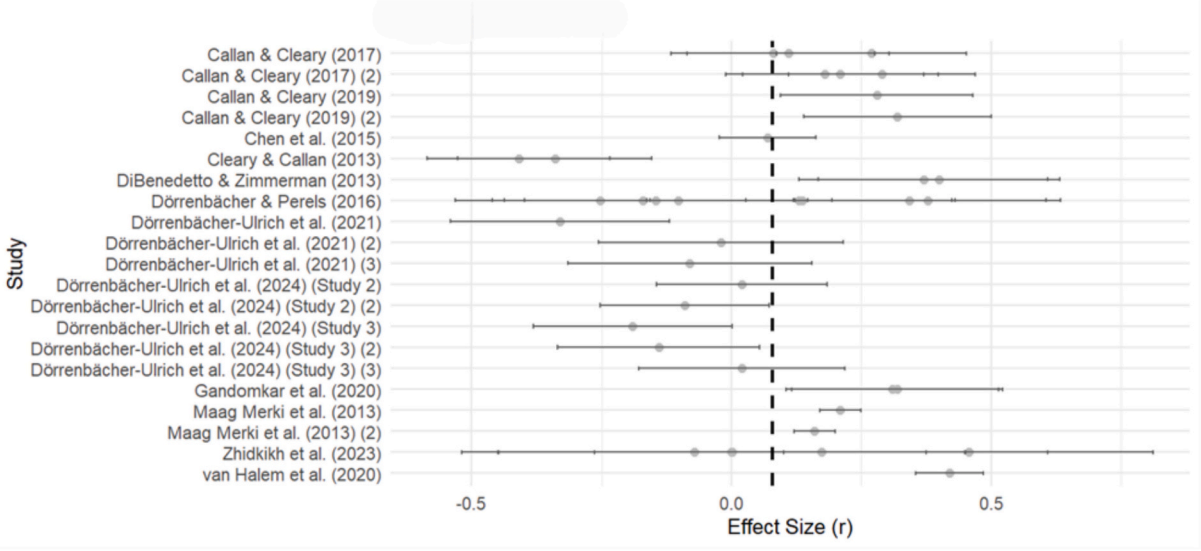


Fig. 3. Forest plot of the cognition component.
Note. Forest plots were created using the original effect sizes from the studies while effect sizes for RVE and moderator analyses were Fisher *Z*-transformed beforehand. For studies by German researchers investigating relationships with grades (which are reverse coded in Germany) we used the absolute value in analyses but the original negative correlations in the forest plots (see Table 3 for which studies are affected).

Table 6
Subgroup analysis of the cognition component.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Self-report questionnaire	8	20	0.07	[−0.16, 0.29]	0.01	0.07
Strategy knowledge test	4	4	0.18	[−0.01, 0.36]	0.01	0.07
Microanalysis	4	8	0.24	[−0.05, 0.50]	0.01	0.07
Behavioural traces	5	8	0.28	[0.14, 0.42]	0.01	0.07

Note. n = number of studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

Table 7
Subgroup analysis of the metacognition component.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Self-report questionnaire	9	21	0.18	[−0.02, 0.38]	0.05	0.23
Strategy knowledge test	4	4	0.16	[0.10, 0.23]	0.05	0.23
Microanalysis	7	26	0.35	[−0.22, 0.74]	0.05	0.23

Note. n = number of studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

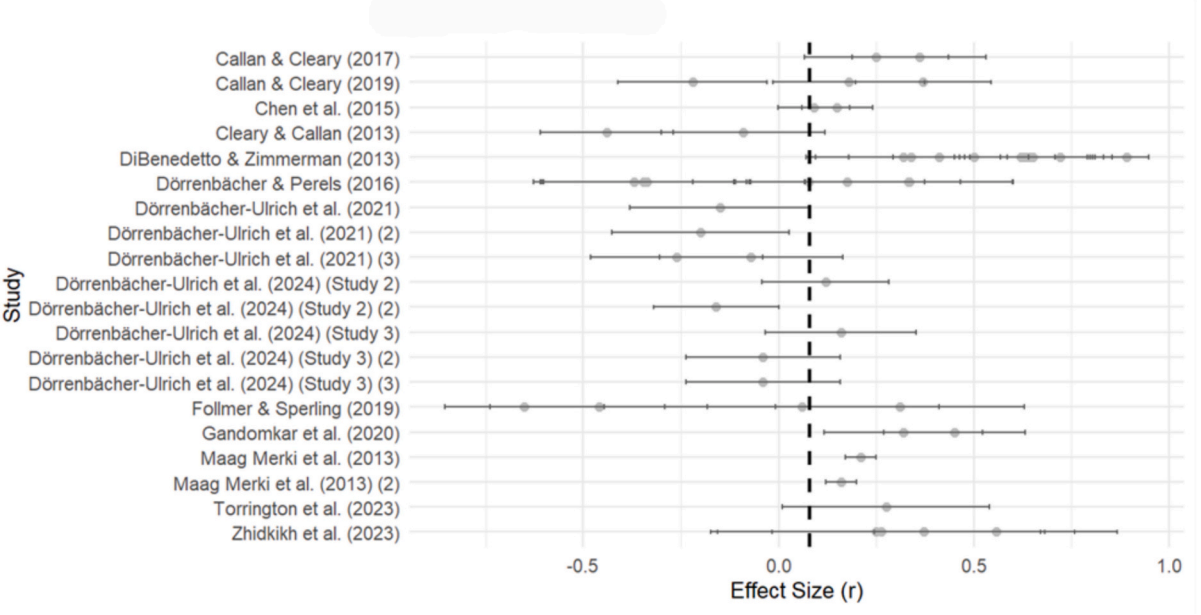


Fig. 4. Forest plot of the metacognition component.
Note. Forest plots were created using the original effect sizes from the studies while effect sizes for RVE and moderator analyses were Fisher *Z*-transformed beforehand. For studies by German researchers investigating relationships with grades (which are reverse coded in Germany) we used the absolute value in analyses but the original negative correlations in the forest plots (see Table 3 for which studies are affected).

effect size for strategy knowledge tests was significantly different from zero ($r = 0.16$, $t(2.97) = 7.74$, $p < .01$). This means that strategy knowledge tests significantly predict the outcome variable. The reason the other two moderators did not differ from zero might be due to large variability in the effect sizes. The reason moderators did not statistically differ from each other might be due to collinearity (Lipsey, 2003).

For the component *motivation*, this led to the inclusion of 7 articles with $n = 14$ studies and $k = 26$ effect sizes. The forest plot is presented in Fig. 5. The robust variance estimations *t*-test was significant ($t(12.9) = 4.13$, $p < .01$). Overall effect size was $r = 0.21$ ($SE = 0.05$, 95 % CI [0.10, 0.32]). Statistics for heterogeneity were $I^2 = 86.78$ and $\tau^2 = 0.05$.

Moderator analyses were carried out for the following instrument subgroups: self-reports, strategy knowledge tests, and microanalysis. Results for each subgroup can be taken from Table 8. Pairwise comparisons were calculated to see whether the subgroups were statistically different from each other.

Overall, results showed medium effect sizes (Cohen, 1992) for self-report questionnaires ($r = 0.28$) and microanalysis ($r = 0.26$). None of the pairwise comparisons were statistically significant. The effect size for self-report questionnaires was statistically different from zero ($r = 0.28$, $t(3.23) = 9.55$, $p < .01$). This suggests that self-report questionnaires significantly predict the outcome variable but cannot be statistically distinguished from the other moderators. This might be due to insufficient power or collinearity/overlaps between the moderators (Lipsey, 2003).

3.5. Research question 4: Is age a moderator of the correlation between different SRL instruments?

To answer this research question, moderator analyses using *robumeta* were conducted as well. We used the same data as for research question 1 just with a different cluster variable (age instead of instrument type). Age was operationalised as 1 = primary/middle school, 2 = high school, and 3 = university, as most studies did not report the actual age of the participants but only their level of education. Primary/middle schoolers were in classes 5 to 8. On average, children in (late) primary and middle school are between 10 and 14 years old. High school was everyone from 9th to 12th grade, thus approximately 14 to 18 years old. University students have a broader possible age range but are, on average, somewhere between 18 and 30 years old. Results for each subgroup can be seen in Table 9. Pairwise comparisons were calculated to analyse whether the three moderator levels differed significantly from each other.

Table 8
Subgroup analysis of the motivation component.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Self-report questionnaire	7	15	0.28	[0.20, 0.37]	0.01	0.12
Strategy knowledge test	3	3	0.04	[−0.04, 0.11]	0.01	0.12
Microanalysis	4	8	0.26	[−0.35, 0.72]	0.01	0.12

Note. n = number of studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

Overall, the results showed small to medium effect sizes (Cohen, 1992) for all age groups. In our sample of studies, university students produced the descriptively largest effect size with $r = 0.21$, which means that correlations between any kind of instruments were highest for this population. The differences between subgroups were not statistically significant. Effect sizes for primary/middle school ($r = 0.17$, $t(3.94) = 7.08$, $p < .001$) and university students ($r = 0.21$, $t(3.44) = 3.96$, $p < .01$) are significantly different from zero.

4. Discussion

4.1. Summary of results

The goal of this meta-analysis was to test how different instruments for measuring SRL relate to each other and achievement, whether there are differences between instruments in measuring individual SRL components, and whether age influences the relations between instruments. To determine this, we analysed 20 studies that measured SRL in a

Table 9
Subgroup analysis for research question 4.

Subgroup	*n*	*k*	*r*	95 % *CI*	τ^2	τ
Primary/middle school	10	101	0.17	[0.10, 0.23]	0.02	0.16
High school	6	104	0.08	[−0.09, 0.26]	0.02	0.16
University	12	146	0.21	[0.05, 0.35]	0.02	0.16

Note. n = number of studies per group, k = number of effect sizes per group, r = effect size, *CI* = confidence interval, τ^2 = variance of the true effect, τ = standard deviation of the true effect.

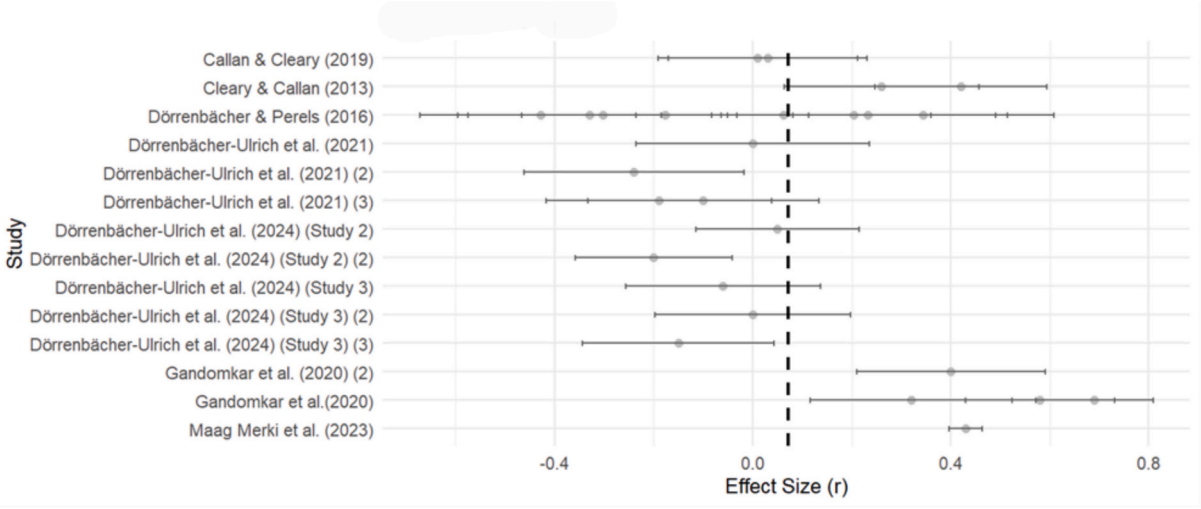


Fig. 5. Forest plot of the motivation component.
Note. Forest plots were created using the original effect sizes from the studies while effect sizes for RVE and moderator analyses were Fisher *Z*-transformed beforehand. For studies by German researchers investigating relationships with grades (which are reverse coded in Germany) we used the absolute value in analyses but the original negative correlations in the forest plots (see Table 3 for which studies are affected).

multimethod way in samples of students. We compared different possible instrument combinations to each other, either two offline, two online, or a mix of online and offline instruments. Compared with mixed instrument combinations, we expected the same types of instruments to have higher correlations (Callan & Cleary, 2018). This expectation was partially confirmed by our data. Moderator analyses suggest that combining two online instruments produces the descriptively highest effect sizes ($r = 0.24$) followed by a combination of mixed methods ($r = 0.15$). The effect sizes do not significantly differ from each other. We further hypothesised that online instruments would perform better in predicting study-specific achievement, such as tests, whereas offline instruments would be better at predicting general academic achievement, such as grades (Callan & Cleary, 2018). This hypothesis was not confirmed by our data. Descriptively, the highest effect size was between online measures and general academic achievement ($r = 0.25$), closely followed by offline measures and study-specific achievement ($r = 0.24$) – which is the opposite of Callan & Cleary's theory. Both effect sizes, as well as the effect size for offline measures and general academic achievement, are statistically different from zero but not from each other. This suggests that the moderators do significantly explain variance in the effect size but cannot be statistically distinguished from each other, which might be due to reasons such as collinearity or small statistical power (Lipsey, 2003). Additionally, we hypothesised that certain instrument types would be better at measuring specific SRL components. Thus, we examined correlations between SRL instruments and achievement for cognition, metacognition, and motivation separately. These analyses are only meant as an approximation to the research question as it is highly complex and it is doubtful that this approach can fully answer it. We assumed that cognition may best be measured using strategy knowledge tests, metacognition using microanalysis, and motivation using self-report questionnaires (Dörrenbächer-Ulrich et al., 2021). While we got significant results overall when conducting RVE, comparisons between subgroups are not statistically significant for any of the components. Nevertheless, for each component there is at least one subgroup that statistically differs from zero ($p < .05$). According to our analyses, and by looking at the descriptively highest effect size, cognition may be best measured using behavioural traces ($r = 0.28$), metacognition using microanalysis ($r = 0.35$), and motivation using self-report questionnaires ($r = 0.28$). None of the pairwise comparisons between the moderators for any component are statistically significant but for cognition behavioural traces are significantly different from zero, and for the component of motivation self-report questionnaires are significantly different from zero as well. For metacognition, microanalysis is not significantly different from zero, which might be due to the high variability in effect sizes. Finally, we tested whether age is a moderator for the relationship between different instruments. Our results show the descriptively highest effect size for university students ($r = 0.21$) which is also significantly different from zero. This means that correlations between any type of instruments seem to be highest in this population. Pairwise comparisons with the effect sizes of primary and high school students were not statistically significant.

4.2. Discussion of results

While overall effect sizes calculated using RVE are all statistically significant, we could only find a few significant results when it comes to moderator analyses. These also only partially confirm our hypotheses. Results of the first research question would suggest that combining two online instruments produces higher correlations than using two offline instruments or a combination of both. This is not an unexpected result because past empirical research already suggests that pairing the same type of instruments produces higher correlations, whereas combining the two results in only marginal correlations if any (Callan & Cleary, 2018). Most articles have examined a combination of microanalysis and behavioural traces (e.g., Callan & Cleary, 2019; Dörrenbächer-Ulrich et al., 2021; Gandomkar et al., 2020) which suggests that combining these two leads to more reliable results or that they might complement each other. Against our expectations, combining mixed instruments produces similar effect sizes as combining two offline instruments ($r = 0.15$ vs. $r = 0.14$). There were also $k = 202$ effect sizes taken into account for the mixed subgroup while the other two subgroups only included around seventy effect sizes each. This indicates that most studies that measure SRL with more than one instrument follow recommendations of combining different measurement methods (Perry & Rahim, 2011) and achieve small correlative results. Differences between the three groups were also not statistically significant. However, the effect sizes of the offline/offline and online/offline group are significantly different from zero, while the effect size of the online/online group, while descriptively the largest, does not significantly differ from zero. This might be due to the sample size and the wide range of included correlations in this group.

Regarding achievement, our results are the opposite of the theory suggested by Callan and Cleary (2018), which states that online measures have larger correlations with study-specific achievement and offline measures with general academic achievement. Descriptively we find the highest effect size between online measures and general academic achievement ($r = 0.25$) followed by offline measures and study-specific achievement ($r = 0.24$). Mean effect size for online + study is close to zero while mean effect size for offline + academic is $r = 0.20$. Comparisons between the groups are not statistically significant. However, the effect sizes for all comparisons except online + study significantly differ from zero. As the results concerning SRL instruments and achievements reported in past research are mixed, this result is not that contrarian. Some studies found both instrument types to be good predictors of achievement, with no indication of one being better for certain achievement variables than the other (Dörrenbächer-Ulrich et al., 2021; Gandomkar et al., 2020). Other studies found online (Cleary et al., 2015; DiBenedetto & Zimmerman, 2013) or offline measures (Cleary & Chen, 2009; Jamieson-Noel & Winne, 2003) superior for any performance. The variety of online and offline instruments as well as performance measures used in each study is also too large to draw definitive conclusions regarding the relationship between different SRL instruments and performance outcome variables. Other factors might influence the relationship between SRL instruments and achievement as well. Variables like cognitive load could have an effect, with online instruments potentially causing too much extraneous load and therefore lowering study-specific performance later on (Sweller, 2011; Wirth et al., 2020), which might explain the null correlation in our results. Concerning offline instruments and academic performance the Dunning-Kruger-Effect or just low metacognitive abilities might play a role (Kruger & Dunning, 1999). Students might be over-assessing their abilities (or just not be accurate in their assessment in either direction), which might explain why the effect size for this group is descriptively a bit lower. It must be noted that the subgroups do not differ, which might be due to low statistical power or collinearity (Lipsey, 2003).

Regarding the assessment of different components, correlations between instrument and performance for the component of cognition are descriptively highest for behavioural traces ($r = 0.28$), for metacognition for microanalysis ($r = 0.35$), and for motivation for self-report measures ($r = 0.28$). Comparisons between instrument subgroups did not yield any significant results for any component. Even then, the suggested direction of these results is reasonable, as behavioural traces are used to measure the processing of learning material, which is where cognition is crucial (e.g., Callan & Cleary, 2018). Furthermore, behavioural traces have been used in many studies. Processes such as viewing learning material, accessing quizzes or assignments, or searching for more information in the learning environment (Du et al., 2023) are all processes that can be tracked and analysed using trace data. Behavioural traces take into account whether students highlight words, draw diagrams, or maybe leave pages blank (Gandomkar et al., 2020), which can provide information about which learning strategies students know and use, as well as their overall cognitive abilities. Thus, behavioural traces

are a well-established and validated instrument and are expected to produce reliable results for the component they intend to measure (Rovers et al., 2019). Although microanalysis measures all three components, it primarily focuses on metacognition. Compared with the other instruments in the studies, it was mainly used to assess metacognition, and in some studies, it was used to assess exclusively metacognition (e.g., Callan & Cleary, 2018). Additionally, it is the only instrument capable of assessing metacognition during the task instead of retrospectively, as in self-report questionnaires (Perels et al., 2020). Retrospective assessment of metacognition is bound to come with a loss of data, as the process of "thinking about thinking" is very context-specific and, therefore, best assessed during the specific task in real time (Dörrenbächer-Ulrich et al., 2021). Nevertheless, the descriptively large effect size for microanalysis that we found in our data is not significantly different from zero ($p > .05$). There might be too much variability in the included effect sizes. Nearly the opposite of what can be said about metacognition can be said for motivation, so we assumed it would be best assessed using self-report questionnaires. If long-term intrinsic motivation is considered the best for successful self-regulated learning, as Zimmerman (2000) does in his model, assessing it with a questionnaire that can consider multiple learning situations from the past, present, and future simultaneously seems to be the most logical procedure. Other assessment methods either only focus on short-term motivation (e.g., microanalysis) or not person-specific motivation (e.g., strategy knowledge tests) and might, therefore, not produce the best results. In our analyses, self-report questionnaires showed an effect size of $r = 0.28$, making it the descriptively largest effect size, which was also significantly different from zero ($p < .05$). Pairwise comparisons did not find any significant differences between the three subgroups suggesting that while self-report do explain variance in the data, they cannot be statistically distinguished from the other moderators. This might be due to statistical power as we also must consider the reduced sample size when analysing this component, as we could only include $n = 7$ articles with $k = 26$ effect sizes. This small sample size also shows that many articles didn't include motivational factors in their definition or examination of SRL. Indeed, it might be that motivation is oftentimes separated from the other SRL components and investigated on its own, with a research field focusing entirely on the motivational factors of SRL (e.g., Pintrich, 2004; Schunk & Zimmerman, 2008). Microanalysis, as the second highest albeit non-significant effect size, can also provide valuable insights into motivation, e.g., by focusing on self-reflection processes after a task (Cleary et al., 2021). Many studies using microanalysis to measure motivation also take attributional processes into account, which are not only important for the learning situation at hand but can also influence later learning (Cleary et al., 2021). Microanalysis can therefore make valuable statements about a student's motivation, which can serve as a guidance to influence general motivation and learning outcomes. It should also be kept in mind that the way we operationalised the research question might not lead to its exact answer. As we correlated instruments with performance outcomes for each component, it might be that other factors influence the effect sizes and not the use of the instrument itself. For example, cognitive load theory might play a role here as well. Research has found a bi-directional relationship between SRL and cognitive load, where high cognitive load can impair SRL, while effective SRL strategies can help with managing cognitive load and improving learning outcomes (Seufert, 2020). It could be that online instruments such as microanalysis and behavioural traces stimulate the usage of effective SRL strategies, which in turn could lead to positive relations with learning outcomes. Meanwhile, offline instruments might introduce extraneous load due to their design (e.g., by being too long; Johnson et al., 1990) which could in turn have an adverse effect on learning outcomes.

Correlations between different SRL instruments seem to be higher for the subgroup of university students ($r = 0.21$), at least descriptively. This may be because different age groups require different assessment methods (Perels et al., 2020). Young children, such as preschool or elementary school students, cannot complete the most commonly used SRL measure (e.g., self-report questionnaires) and, therefore, require alternative methods. Many of those methods are in the early stages of development and not as qualitatively sound as the more established assessments (e.g., Moreira et al., 2022). Furthermore, metacognitive skills, crucial for self-regulation and behaviour, may differ between age groups (Grainger et al., 2016). Given that metacognition is important for remembering past events and imagining one's behaviour in the future, which assessment methods such as questionnaires require, metacognition is essential to complete certain measures (Grainger et al., 2016). It is assumed that metacognitive skills emerge between the ages of eight and ten and then take some time until they fully develop (Veenman et al., 2006). Furthermore, metacognition should be fully developed by the time someone starts university but may not be as pronounced in school-aged children, which may offer another explanation as to why relations between instruments are higher for university students than they are for primary, middle, and high schoolers.

Overall, finding only small or medium correlative results between different SRL measures is not bad. On the contrary, SRL is considered a multi-faceted construct that would be hard to measure using only one instrument. Different measures target different aspects of SRL; some could potentially over- or underrepresent a student's "real" SRL (Callan & Cleary, 2018), leading to poor correlative results between the instruments. This, once again, is why a multimethod approach is so important: Only by combining various instruments can one get a fuller picture of a person's SRL.

4.3. Limitations

One limitation is the number of studies available for analysis. The low sample size is insofar problematic for the results and statistical analyses as that many procedures that can be used for meta-analysis require a certain number of studies to function and produce reliable results (Brockwell & Gordon, 2001). Meta-regression, for example, which we used to find subgroup differences, is often underpowered (Borenstein et al., 2009). Especially models with fewer than 10 studies lead to low power and high Type I error rates (Fu et al., 2011). Most authors suggest at least ten studies per covariate (Thompson & Higgins, 2002), which was not the case in our data. On the other hand, meta-regressions work well with dependent effect sizes and in combination with robust variance estimations. Considering that alternatives such as subgroup analyses assume independent effect sizes (Harrer et al., 2021) and we had used RVE already, we decided to still go ahead and do meta-regressions. As most of our meta-regressive results are not significant while results for overall effect sizes calculated using RVE are, it might be that the statistical power was not high enough due to the sample size to find a significant effect. The main reason so few studies were retrieved is that multimethod SRL assessment is still rare, at least in the sense of using two or more different SRL instruments. We found a handful of more studies that e.g., used multiple questionnaires to assess SRL (e.g., Moote et al., 2013; Mountain et al., 2023), but this did not qualify as multimethod SRL assessment by our definition. Another reason is the strict inclusion and exclusion criteria. The most significant limitation was the required inclusion of at least two SRL components. We deemed this necessary, as SRL is a process defined by its three distinct components, which we wanted to depict as best as possible. However, many SRL studies focus on only one component at a time, mostly metacognition (e.g., Veenman & Van Cleef, 2019). We felt that including only studies focusing on metacognition would not do SRL justice, as the construct is much more multifaceted.

Regarding the statistical analyses and interpretation of results, the high heterogeneity in our data should also be mentioned. While a certain amount of heterogeneity is good and can provide insights (Borenstein et al., 2019), there appears to be a lot of variance in our effect sizes as I^2 statistics were consistently above 80 % (Higgins et al., 2003). To control for this, we used random effects models on one hand, as they take

within- and between-study variance into account and provide a more accurate estimate for when heterogeneity exists (Borenstein et al., 2019). On the other hand, we conducted moderator analyses to figure out what potential study characteristics causing variance might be (Borenstein et al., 2019). While the methods applied should help with figuring sources of heterogeneity out, we can also see that some of our proposed moderators still contain high heterogeneity (see τ and τ^2 for research question 3). This suggests that there are differences between studies that we have not considered. This is an important limitation when interpretating these results as unexplained variance make the interpretation more complex and less reliable. Further, as pairwise comparisons between all moderators were statistically nonsignificant, it might also be that moderators have high collinearity and overlap (Lipsey, 2003). This might be because some groups of researchers were over-represented in our data, which also led them to be over-represented in the subgroups. This is a problem insofar as researchers tend to use the same or at least similar instruments across their publications, which would make the data in a subgroup more similar and can, therefore, lead to bias.

One more thing worth mentioning is that when it comes to interpreting the results of SRL-instruments, there are many ways to go about it. Especially online instruments such as trace data or microanalysis allow for the use of many different indicators (e.g., Du et al., 2023) – with some being more and less valid measures of high-quality SRL. When analysing our data, we had to aggregate all the different indicators used by different researchers together to represent one category of SRL-measurement, e.g., "trace data". When conducting a meta-analysis, this cannot be avoided but, of course, introduces another source of heterogeneity in the data.

Another limitation concerns the potential presence of publication bias. We used two different tests to determine whether there might be publication bias in our data (Egger's and PET-PEESE), which resulted in differing results. The Egger's test showed publication bias (in the form of small study bias) in our data, while the PET test didn't. We used two different tests because it has been suggested that the Egger's test does not produce reliable results in the case of dependent effect sizes (Park et al., 2025). Further, the logical countermeasure to finding publication bias using the Egger's test would be trim-and-fill-procedures – which does not produce valid results in combination with RVE (Peters et al., 2007). Therefore, we added the PET-PEESE test, which is compatible with RVE (Stanley & Doucouliagos, 2014) and has an added correction (PEESE) in case of publication bias. The PET-PEESE method is known to perform poorly in the case of small samples (Stanley, 2017). Overall, we cannot be sure about the existence or non-existence of publication bias in our data, so the results should be interpreted cautiously.

It also has to be mentioned that we only included one rater for the coding and extraction part of the data. This also represents a significant limitation as PRISMA standards recommend at least two (Page et al., 2021).

4.4. Implications

The main conclusion from the results is that different types of SRL instruments relate to each other and can be used in different combinations. This is an important result, reinforcing the demand that researchers use a multimethod approach when conducting studies on SRL. Moreover, the results show that multimethod SRL research continues to be insufficient. This is concerning, considering multimethod research has been demanded for years, but this call only seems to have been answered by a handful of researchers. This is increasingly evident given that multiple studies included in our analyses have been conducted by the same researchers, whereas other SRL research groups are not represented at all. Hopefully, this meta-analysis will serve as a reminder and encourage researchers to investigate multimethod SRL assessment further.

In terms of practical implications, our results may encourage teachers and other practitioners to assess their students' SRL more widely. The importance of SRL for education is indisputable but is often not considered in practice. The knowledge that various instruments measuring SRL exist and produce sufficiently high correlations with achievement may remove the barrier of "picking the right one" and encourage practitioners to experiment with a variety of instruments or use the ones best suited for them and their needs. In particular, practitioners or school psychologists could use different SRL instruments to identify different problem fields for individual students. For example, they could use behavioural traces to determine which specific strategies a student uses and how they are implemented during a particular task. Combining this data with performance data such as tests could also show whether the student's learning works and produces the desired results. Microanalysis could be used to evaluate how a student prepares for task, how they set goals, and how they reflect. This knowledge can ultimately help shape students' goal-setting strategies and attributional processes – which in turn can influence performance in a broader educational context.

4.5. Conclusion

In conclusion, our results suggest that multimethod assessment is important and sensible and a clear path for going forward in the future. Much more research is required to answer the research questions we posed satisfactorily. In particular, more studies are necessary to determine how populations of different ages react to different SRL assessment types and whether, in turn, there are differences in the prediction of achievement. Furthermore, for future meta-analyses, it would be interesting to include multimethod studies focusing only on one SRL component. This should lead to more relevant articles that could more reliably answer the question of which instrument best assesses which component. Knowing which assessment method is best for which component is valuable for making SRL assessment as precise as possible.

Open science

This meta-analysis was preregistered: https://osf.io/2vjgs/?view_only=0a6978e8818145cfa2670eda0cc0caed.

Data and analysis scripts can also be found here: https://osf.io/s57y4/?view_only=8f1de8f28cb44b009811e2dfc96c8b49.

CRediT authorship contribution statement

Julia Ruhl: Writing – original draft, Formal analysis, Conceptualization, Writing – review & editing, Methodology, Data curation. **Franziska Perels:** Supervision. **Laura Dörrenbächer-Ulrich:** Conceptualization, Supervision.

Ethics statement

The meta-analysis complies with all ethical standards.

Funding

This study was funded by the German Research Foundation (DO 2173/2-1).

Declaration of competing interest

We have no conflict of interest to disclose.

References*

Anderson, H., Coltman, P., Page, C., & Whitebread, D. (2003, August). Developing independent learning in children aged 3–5. In *European Association for research on learning and instruction 10th biennial conference, Padova, August*.

*Anthony, J. S., Clayton, K. E., & Zusho, A. (2013). An investigation of students' self-regulated learning strategies: Students' qualitative and quantitative accounts of their learning strategies. *Journal of Cognitive Education and Psychology, 12*(3), 359–373. https://doi.org/10.1891/1945-8959.12.3.359

Artelt, C. (2000). *Strategisches Lernen [Strategic Learning]*. Münster: Waxmann.

Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist, 45*(4), 210–223. https://doi.org/10.1080/00461520.2010.515934

*Bellhäuser, H., Dignath, C., & Theobald, M. (2023). Daily automated feedback enhances self-regulated learning: A longitudinal randomized field experiment. *Frontiers in Psychology, 14*, Article 1125873. https://doi.org/10.3389/fpsyg.2023.1125873

Benick, M., Dörrenbächer-Ulrich, L., & Perels, F. (2018). Process evaluation of differential effects within an intervention to improve self-regulated learning towards the end of primary school. *Unterrichtswissenschaft, 46*, 379–407.

Benick, M., Dörrenbächer-Ulrich, L., Weißenfels, M., & Perels, F. (2021). Fostering self-regulated learning in primary school students: Can additional teacher training enhance the effectiveness of an intervention? *Psychology Learning & Teaching, 20*(3), 324–347. https://doi.org/10.1177/14757257211013638

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*(6), 445–457. https://doi.org/10.1016/S0883-0355(99)00014-2

Boekaerts, M. (2006). Self-regulation and effort investment. In W. Damon, R. A. Lerner R, K. A. Renninger, & I. E. Sigel (Eds.), *Handbook of child psychology volume 4: Child psychology in practice* (pp. 345–377). New York, NY: Wiley.

Boekaerts, M. (2007). Understanding students' affective processes in the classroom. In B. Schulz, & R. E. Pekrun (Eds.), *Emotion in education* (pp. 37–56). San Diego, CA: Academic Press.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199–231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Borenstein, M., Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2019). Heterogeneity in meta-analysis. *The handbook of research synthesis and meta-analysis, 3*, 453–470.

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

Brockwell, S., & Gordon, I. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine, 20*(6), 825–840. https://doi.org/10.1002/sim.650

Brophy, J. (2005). Goal theorists should move on from performance goals. *Educational Psychologist, 40*(3), 167–176. https://doi.org/10.1207/s15326985ep4003_3

*Callan, G. L., & Cleary, T. J. (2018). Multidimensional assessment of self-regulated learning with middle school math students. *School Psychology Quarterly, 33*(1), 103–111. https://doi.org/10.1037/spq0000198

*Callan, G. L., & Cleary, T. J. (2019). Examining cyclical phase relations and predictive influences of self-regulated learning processes on mathematics task performance. *Metacognition and Learning, 14*(1), 43–63. https://doi.org/10.1007/s11409-019-09191-x

Chang, C. C., Tseng, K. H., Liang, C., & Chen, T. Y. (2013). Using e-portfolios to facilitate university students' knowledge management performance: E-portfolio vs. non-portfolio. *Computers & Education, 69*, 216–224. https://doi.org/10.1016/j.compedu.2013.07.017

*Chen, P. P., Cleary, T. J., & Lui, A. M. (2015). Examining parents' ratings of middle-school students' academic self-regulation using principal axis factoring analysis. *School Psychology Quarterly, 30*(3), 385–397. https://doi.org/10.1037/spq0000098

Cleary, T. J. (2011). Emergence of self-regulated learning microanalysis: Historical overview, essential features, and implications for research and practice. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 329–345). New York, NY: Routledge.

*Cleary, T. J., & Callan, G. L. (2013). Student self-regulated learning in an urban high school: Predictive validity and relations between teacher ratings and student self-reports. *Journal of Psychoeducational Assessment, 32*(4), 295–305. https://doi.org/10.1177/0734282913507

Cleary, T. J., & Callan, G. L. (2018). Assessing self-regulated learning using microanalytic methods. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 338–351). New York: Routledge.

*Cleary, T. J., Callan, G. L., Malatesta, J., & Adams, T. (2015). Examining the level of convergence among self-regulated learning microanalytic processes, achievement, and a self-report questionnaire. *Journal of Psychoeducational Assessment, 33*(5), 439–450. https://doi.org/10.1177/07342829155947

Cleary, T. J., Callan, G. L., & Zimmerman, B. J. (2012). Assessing self-regulation as a cyclical, context-specific phenomenon: Overview and analysis of SRL microanalytic protocols. *Education Research International.*. https://doi.org/10.1155/2012/428639

Cleary, T. J., & Chen, P. P. (2009). Self-regulation, motivation, and math achievement in middle school: Variations across grade level and math context. *Journal of School Psychology, 47*(5), 291–314. https://doi.org/10.1016/j.jsp.2009.04.002

Cleary, T. J., & Kitsantas, A. (2017). Motivation and self-regulated learning influences on middle school mathematics achievement. *School Psychology Review, 46*(1), 88–107. https://doi.org/10.17105/SPR46-1.88-107

Cleary, T. J., & Russo, M. J. (2023). A multilevel framework for assessing self-regulated learning in school contexts: Innovations, challenges, and future directions. *Psychology in the Schools, 61*(1), 80–102. https://doi.org/10.1002/pits.23035

Cleary, T. J., Slemp, J., Reddy, L. A., Alperin, A., Lui, A., Austin, A., & Cedar, T. (2021). Characteristics and uses of SRL microanalysis across diverse contexts, tasks, and populations: A systematic review. *School Psychology Review, 52*(2), 159–179. https://doi.org/10.1080/2372966X.2020.1862627

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools, 41*(5), 537–550. https://doi.org/10.1002/pits.10177

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

*DiBenedetto, M. K., & Zimmerman, B. J. (2013). Construct and predictive validity of microanalytic measures of students' self-regulation of science learning. *Learning and Individual Differences, 26*, 30–41. https://doi.org/10.1016/j.lindif.2013.04.004

Dignath, C., van Ewijk, R., Perels, F., & Fabriz, S. (2023). Let learners monitor the learning content and their learning behavior! A meta-analysis on the effectiveness of tools to foster monitoring. *Educational Psychology Review, 35*(2), 62. https://doi.org/10.1007/s10648-023-09718-4

Dinsmore, D., Alexander, P., & Loughlin, S. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*, 391–409. https://doi.org/10.1007/s10648–008–9083-6

Dörr, L., & Perels, F. (2018). Multiperspektivische Erfassung der Selbstregulationsfähigkeit von Vorschulkindern [A multiperspective approach to assessing preschoolers' self-regulating ability]. *Frühe Bildung, 7*(2), 98–106. https://doi.org/10.1026/2191-9186/a000359

*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

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

**Dörrenbächer-Ulrich, L., Weißenfels, M., Russer, L., & Perels, F. (2021). Multimethod assessment of self-regulated learning in college students: Different methods for different components? *Instructional Science, 49*(1), 137–163. https://doi.org/10.1007/s11251-020-09533-2

Du, J., Hew, K. F., & Liu, L. (2023). What can online traces tell us about students' self-regulated learning? A systematic review of online trace data analysis. *Computers & Education, 201*, Article 104828. https://doi.org/10.1016/j.compedu.2023.104828

Endedijk, M. D., Brekelmans, M., Sleegers, P., & Vermunt, J. D. (2016). Measuring students' self-regulated learning in professional education: Bridging the gap between event and aptitude measurements. *Quality & Quantity, 50*(5), 2141–2164. https://doi.org/10.1007/s11135-015-0255-4

Entwistle, N. (1988). Motivational factors in students approaches to learning. In *Learning strategies and learning styles*. Plenum Press.

EU Council. (2002). *Council resolution of 27 June 2002 on lifelong learning*. CELEX.

Fan, Y., Matcha, W., Uzir, N. A. A., Wang, Q., & Gašević, D. (2021). Learning analytics to reveal links between learning design and self-regulated learning. *International Journal of Artificial Intelligence in Education, 31*(4), 980–1021. https://doi.org/10.1007/s40593-021-00249-z

*Fan, Y., Rakovic, M., van Der Graaf, J., Lim, L., Singh, S., Moore, J., … Gašević, D. (2023). Towards a fuller picture: Triangulation and integration of the measurement of self-regulated learning based on trace and think aloud data. *Journal of Computer Assisted Learning, 39*(4), 1303–1324. https://doi.org/10.1111/jcal.12801

Fan, Y., van der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., … Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. *Metacognition and Learning, 17*(3), 949–987. https://doi.org/10.1007/s11409-022-09291-1

Fisher, Z., & Tipton, E. (2015). *Robumeta: An R-package for robust variance estimation in meta-analysis*.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906.

*Follmer, D. J., & Sperling, R. A. (2019). Examining the role of self-regulated learning microanalysis in the assessment of learners' regulation. *The Journal of Experimental Education, 87*(2), 269–287. https://doi.org/10.1080/00220973.2017.1409184

Fu, R., Vandermeer, B. W., Shamliyan, T., O'Neil, M. E., Yazdi, F., Fox, S., & Shekelle, P. G. (2011). *Handling continuous outcomes in quantitative synthesis. AHRQ Methods Guide for Comparative Effectiveness Reviews*.

*Gandomkar, R., Yazdani, K., Fata, L., Mehrdad, R., Mirzazadeh, A., Jalili, M., & Sandars, J. (2020). Using multiple self-regulated learning measures to understand medical students' biomedical science learning. *Medical Education, 54*(8), 727–737. https://doi.org/10.1111/medu.14079

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

Grassinger, R. (2011). Selbstregulation beim Wechseln der Lernumwelt [self-regulation on changing the learning environment]. In M. Dresel, & L. Lämmle (Eds.), *Motivation, Selbstregulation und Leistungsexzellenz [motivation, self-regulation and achievement excellency]* (pp. 179–197). Münster: LIT-Verlag.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning, 2*, 107–124. https://doi.org/10.1007/s11409-007-9016-7

* Note. References marked with an asterisk indicate studies included in this meta-analysis.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R*. https://doi.org/10.1201/9781003107347

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

Heirweg, S., De Smul, M., Merchie, E., Devos, G., & Van Keer, H. (2020). Mine the process: Investigating the cyclical nature of upper primary school students' self-regulated learning. *Instructional Science, 48*(4), 337–369. https://doi.org/10.1007/s11251-020-09519-0

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.), 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hutchinson, L. R., Perry, N. E., & Shapka, J. D. (2021). Assessing young children's self-regulation in school contexts. *Assessment in Education: Principles, Policy & Practice, 28*(5–6), 545–583. https://doi.org/10.1080/0969594X.2021.1951161

Jacob, L. (2020). *Investigating self-regulated learning in preschoolers (Dissertationsschrift, Universität des Saarlandes)*. Publikationen.sulb.uni-saarland.de. https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/30311/1/Dissertation_Uds_Jacob.pdf.

Jamet, É. (2014). An eye-tracking study of cueing effects in multimedia learning. *Computers in Human Behavior, 32*, 47–53. https://doi.org/10.1016/j.chb.2013.11.013

Jamieson-Noel, D., & Winne, P. H. (2003). Comparing self-reports to traces of studying behavior as representations of students' studying and achievement. *Zeitschrift Für Pädagogische Psychologie, 17*(3/4), 159–172.

Jansen, R. S., van Leeuwen, A., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners' self-regulated learning in Massive Open Online Courses. *Computers & Education, 146*, Article 103771. https://doi.org/10.1016/j.compedu.2019.103771

Johnson, M. D., Lehmann, D. R., & Horne, D. R. (1990). The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing, 7*(1), 35–43. https://doi.org/10.1016/0167-8116(90)90030-Q

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., … Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Kitsantas, A., Winsler, A., & Huie, F. (2008). Self-regulation and ability predictors of academic success during college: A predictive validity study. *Journal of advanced academics, 20*(1), 42–68. https://doi.org/10.4219/jaa-2008-867

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00863

Lau, C., Kitsantas, A., & Miller, A. (2015). Using microanalysis to examine how elementary students self-regulate in math: A case study. *Procedia-Social and Behavioral Sciences, 174*, 2226–2233. https://doi.org/10.1016/j.sbspro.2015.01.879

Lau, K. L. (2012). Construction and validation of a Chinese SRL-based reading instruction questionaire. *Educational Research and Evaluation, 18*(5), 489–509. https://doi.org/10.1080/13803611.2012.689730

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science, 587*(1), 69–81. https://doi.org/10.1177/0002716202250791

Lodewyk, K. R., Winne, P. H., & Jamieson-Noel, D. L. (2009). Implications of task structure on self-regulated learning and achievement. *Educational Psychology, 29*, 1–25. https://doi.org/10.1080/01443410802447023

*Maag Merki, K. M., Ramseier, E., & Karlen, Y. (2013). Reliability and validity analyses of a newly developed test to assess learning strategy knowledge. *Journal of Cognitive Education and Psychology, 12*(3), 391. https://doi.org/10.1891/1945-8959.12.3.391

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*(2), 167–171. https://doi.org/10.1016/j.learninstruc.2009.02.012

McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning, 10*(1), 43–75. https://doi.org/10.1007/s11409-014-9132-0

*Metallidou, P., & Vlachou, A. (2010). Children's self-regulated learning profile in language and mathematics: The role of task value beliefs. *Psychology in the Schools, 47*(8), 776–788. https://doi.org/10.1002/pits.20503

Moote, J. K., Williams, J. M., & Sproule, J. (2013). When students take control: Investigating the impact of the CREST inquiry-based learning program on self-regulated processes and related motivations in young science students. *Journal of Cognitive Education and Psychology, 12*(2), 178–196.

Moreira, J. S., Ferreira, P. C., & Simão, A. M. V. (2022). Dynamic assessment of self-regulated learning in preschool. *Heliyon, 8*(8), Article e10035. https://doi.org/10.1016/j.heliyon.2022.e10035

Mountain, K., Teviotdale, W., Duxbury, J., & Oldroyd, J. (2023). Are they taking action? Accounting undergraduates' engagement with assessment criteria and self-regulation development. *Accounting Education, 32*(1), 34–60. https://doi.org/10.1080/09639284.2022.2030240

Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Computers in Human Behavior, 96*, 223–234. https://doi.org/10.1016/j.chb.2018.06.028

*Paans, C., Molenaar, I., Segers, E., & Verhoeven, L. (2018). Temporal variation in children's self-regulated hypermedia learning. *Computers in Human Behavior, 96*, 246–258. https://doi.org/10.1016/j.chb.2018.04.002

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *bmj, 372*. https://doi.org/10.1136/bmj.n71

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

Park, S., Beretvas, S. N., & Smith, T. E. (2025). Extending egger's regression: Detecting outcome reporting bias in meta-analysis on dependent multiple outcomes. *The Journal of Experimental Education*, 1–39. https://doi.org/10.1080/00220973.2025.2477720

Perels, F., Dörrenbächer-Ulrich, L., Landmann, M., Otto, B., Schnick-Vollmer, K., & Schmitz, B. (2020). Selbstregulation und selbstreguliertes Lernen. In *Pädagogische Psychologie* (pp. 45–66). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61403-7_3.

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). New York: Routledge.

Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review, 18*(3), 211–228. https://doi.org/10.1007/s10648-006-9014-3

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine, 26*(25), 4544–4562. https://doi.org/10.1002/sim.2889

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

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*, 801–813. https://doi.org/10.1177/0013164493053003024

Pustejovsky, J. E., Pekofsky, S., & Zhang, J. (2025). *clubSandwich (Version 0.5.11) [R package]*. https://doi.org/10.32614/CRAN.package.clubSandwich

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169–184. https://doi.org/10.1002/mpr.289

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, 28*(3), 225–250. https://doi.org/10.1007/s11092-015-9229-2

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

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19–30. https://doi.org/10.1037/1082-989x.13.1.19

Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 1–15). Routledge.

Schunk, D. H., & Zimmerman, B. J. (2008). Motivation and self-regulated learning. In *Theory, research and applications*.

Seufert, T. (2020). Building bridges between self-regulation and cognitive load—An invitation for a broad and differentiated attempt. *Educational Psychology Review, 32*(4), 1151–1162. https://doi.org/10.1007/s10648-020-09574-6

Shuy, T. (2010). *Self-regulated learning*. TEAL Center. https://lincs.ed.gov/state-resources/federal-initiatives/teal/guide/selfregulated.

Siadaty, M., Gasevic, D., & Hatala, M. (2016). Trace-based micro-analytic measurement of self-regulated learning processes. *Journal of Learning Analytics, 3*(1), 183–214. https://doi.org/10.18608/jla.2016.31.11

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin, 137*(3), 421–442. https://doi.org/10.1037/a0022777

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation, 10*(2), 117–139. https://doi.org/10.1076/edre.10.2.117.27905

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*(5), 581–591.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78.

Sweller, J. (2011). Cognitive load theory. In *, 55. Psychology of learning and motivation* (pp. 37–76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology, 66*, Article 101976. https://doi.org/10.1016/j.cedpsych.2021.101976

Theobald, M., Bäulke, L., Bellhäuser, H., Breitwieser, J., Mattes, B., Brod, G., & Nückles, M. (2023). A multi-study examination of intra-individual feedback loops between competence and value beliefs, procrastination, and goal achievement. *Contemporary Educational Psychology, 74*, Article 102208. https://doi.org/10.1016/j.cedpsych.2023.102208

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

Throndsen, I. (2011). Self-regulated learning of basic arithmetic skills: A longitudinal study. *British Journal of Educational Psychology, 81*(4), 558–578. https://doi.org/10.1348/2044-8279.002008

Tinajero, C., Mayo, M. E., Villar, E., & Martínez-López, Z. (2024). Classic and modern models of self-regulated learning: Integrative and componential analysis. *Frontiers in Psychology, 15*, Article 1307574. https://doi.org/10.3389/fpsyg.2024.1307574

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375. https://doi.org/10.1037/met0000011

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/10.3102/10769986156060

*Torrington, J., Bower, M., & Burns, E. C. (2023). Elementary students' self-regulation in computer-based learning environments: How do self-report measures, observations and teacher rating relate to task performance? *British Journal of Educational Technology, 55*(1), 231–258. https://doi.org/10.1111/bjet.13338

Van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In *Springer international handbooks of education* (pp. 143–156). https://doi.org/10.1007/978-1-4419-5546-3_10

*van Halem, N., Van Klaveren, C., Drachsler, H., Schmitz, M., & Cornelisz, I. (2020). Tracking patterns in self-regulated learning using students' self-reports and online trace data. *Frontline Learning Research, 8*(3), 140–163. https://doi.org/10.14786/flr.v8i3.497

van Hout-Wolters, B. (2000). Assessing active self-directed learning. In P. R. J. Simons, J. van der Linden, & T. Duffy (Eds.), *New learning* (pp. 83–101). Dordrecht: Kluwer.

Veenman, M. V. (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., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology, 73*(3), 357–372. https://doi.org/10.1348/000709903322275885

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. https://hdl.handle.net/11245/1.266074.

Veenman, M. V. J. (2013). Assessing metacognitive skills in computerized learning environments. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies*. New York: Springer.

Veenman, M. V. J., & Van Cleef, D. (2019). Measuring metacognitive skills for mathematics: Students' self-reports versus on-line assessment methods. *ZDM - Mathematics Education, 51*(4), 691–701. https://doi.org/10.1007/s11858-018-1006-5

Viechtbauer, W. (2021). *Metafor: Meta-analysis package for R*.

Winne, P. (2011). A cognitive and metacognitive analysis of self-regulated learning. In D. H. Schunk, & B. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–32). Routledge.

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

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*(4), 551–572. https://doi.org/10.1016/s0361-476x(02)00006-1

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Academic Press. https://doi.org/10.1016/B978-012109890-2/50045-7.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Journal of Psychology, 216*(2), 102–110. https://doi.org/10.1027/0044-3409.216.2.102

Wirth, J., Stebner, F., Trypke, M., Schuster, C., & Leutner, D. (2020). An interactive layers model of self-regulated learning and cognitive load. *Educational Psychology Review, 32*(4), 1127–1149. https://doi.org/10.1007/s10648-020-09568-4

Wolters, C. A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist, 38*(4), 189–205. https://doi.org/10.1207/S15326985EP3804_1

Wolters, C. A., & Won, S. (2018). Validity and the use of self-report questionnaires to assess self-regulated learning. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 307–322). New York: Routledge.

Zeidner, M., & Stoeger, H. (2019). Self-Regulated Learning (SRL): A guide for the perplexed. *High Ability Studies, 30*(1–2), 9–51. https://doi.org/10.1080/13598139.2019.1589369

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review, 38*, 55–68. https://doi.org/10.1016/j.dr.2015.07.001

*Zhidkikh, D., Saarela, M., & Kärkkäinen, T. (2023). Measuring self-regulated learning in a junior high school mathematics classroom: Combining aptitude and event measures in digital learning materials. *Journal of Computer Assisted Learning, 39*(6), 1834–1851. https://doi.org/10.1111/jcal.12842

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction, 22*(6), 413–419. https://doi.org/10.1016/j.learninstruc.2012.03.004

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology, 11*(4), 307–313. https://doi.org/10.1016/0361-476X(86)90027-5

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic press. https://doi.org/10.1016/B978-012109890-2/50031-7.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. https://doi.org/10.3102/0002831207312909

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and an overview. In D. H. Schunk, & B. J. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–26). Routledge.