Educational Psychology Review (2020) 32:979–1002
https://doi.org/10.1007/s10648-020-09569-3

REVIEW ARTICLE

The Relation Between Students' Effort and Monitoring Judgments During Learning: A Meta-analysis

Martine Baars[1] · Lisette Wijnia[2,3] · Anique de Bruin[4] · Fred Paas[1,5]

Published online: 9 September 2020
© The Author(s) 2020

Abstract
Research has shown a bi-directional association between the (perceived) amount of invested effort to learn or retrieve information (e.g., time, mental effort) and metacognitive monitoring judgments. The direction of this association likely depends on how learners allocate their effort. In self-paced learning, effort allocation is usually data driven, where the ease of memorizing is used as a cue, resulting in a negative correlation between effort and monitoring judgments. Effort allocation is goal driven when it is strategically invested (e.g., based on the importance of items or time pressure) and likely results in a positive correlation. The current study used a meta-analytic approach to synthesize the results from several studies on the relationship between effort and monitoring judgments. The results showed that there was a negative association between effort and monitoring judgments ($r = -.355$). Furthermore, an exploration of possible moderators of this association between effort and monitoring was made. The negative association was no longer significant when goal-driven regulation was manipulated. Furthermore, it was found that the type of monitoring judgment (i.e., a weaker association for prospective judgments) and type of task (stronger association for problem-solving tasks relative to paired associates) moderated the relation between effort and monitoring. These results have important implications for future research on the use of effort as a cue for monitoring in self-regulated learning.

Keywords Effort · Monitoring · Cue utilization · Meta-analysis · Metacognitive judgments

Monitoring is a central concept in most models of self-regulated learning (SRL, for an overview of SRL models, see Panadero, 2017). According to the cue utilization perspective (Koriat 1997), monitoring judgments are inferential. That is, learners use a variety of cues to estimate the probability that they will remember or recognize specific knowledge or procedures later on a test. Examples of these cues are learners' beliefs about their memory and how

Martine Baars and Lisette Wijnia contributed equally to this work.

✉ Martine Baars
baars@essb.eur.nl

Extended author information available on the last page of the article

confident they are about remembering learning materials. Next, important aspects of the study situation, like the number of trials, the type of encoding strategies, and the type of test learners expect, can be used as cues. Also, previous task-specific experiences and the perceived relative difficulty of the items could be potential cues (Koriat 1997). Research has shown an association between effort and monitoring judgments which, in line with the cue utilization perspective, suggests that the perceived amount of invested effort in the learning task is being used as a cue to make monitoring judgments (Koriat and Ma'ayan 2005; Koriat et al., 2014b; Undorf and Erdfelder 2011). For example, several studies by Baars and colleagues showed that effort ratings were negatively correlated to monitoring judgments in primary and secondary education when learning to solve problems (Baars et al., 2018; Baars et al., 2014; Baars et al., 2013). Yet, it is still largely unclear what and how aspects of the learning process affect the association between effort and monitoring. For example, if and how the type of monitoring judgment or the type of effort measurement could affect the association between effort and monitoring. Therefore, the current study aimed to assess the association between effort and monitoring judgments made by students in the context of studies on learning and performance and investigated possible moderators in this association.

Data-Driven and Goal-Driven Self-Regulation

Processing fluency, such as study or response time, during encoding, and retrieving, seems to be an important cue for learners' metacognitive judgments. Koriat et al. (2006) proposed a memorizing effort heuristic. According to this heuristic, learners use invested memorizing effort (e.g., mental effort, study time) as a cue. They believe that they will more likely remember easily learned items than items that require more study effort. Subsequently, this belief results in a negative correlation between effort and monitoring judgments, that is, with increasing effort, the monitoring judgments tend to decrease from more confident to less confident of being able to recall or understand the materials (e.g., Koriat et al., 2009a, 2009b; Undorf and Erdfelder 2011). Following this view, metacognitive monitoring judgments are *data driven*. The time needed to encode or retrieve information or the mental effort invested is taken, retrospectively, as a cue for the learners' competence and mastery of the learning material (Koriat et al. 2006; Schneider and Löffler, 2016). This process has also been referred to as the "control affects monitoring" (CM) model (Koriat et al. 2006, 2009a, 2009b). Data-driven self-regulation often takes place during self-paced learning, where learners can spend as much time or effort on learning the material as needed. Therefore, the ease of memorizing can be used as a cue (e.g., Koriat et al., 2014a; Undorf and Erdfelder 2011).

Evidence for the memorizing effort heuristic and how effort is used to make monitoring judgments has been found for word pairs (e.g., Undorf and Erdfelder 2011) as well as problem-solving tasks (Ackerman and Zalmanov 2012). The memorizing effort heuristic was found in children and adults (Koriat et al. 2009a; Koriat et al., 2014a). However, research also suggests age-related improvements in cue utilization, because the negative correlation was found to be weaker or nonsignificant for 7-to-8-year old children compared with older children (Hoffmann-Biencourt et al., 2010; Koriat et al. 2009a).

Research by Koriat and colleagues has shown that the correlation between effort and monitoring judgments is not always negative (Koriat 2008; Koriat et al. 2006; Koriat et al., 2014b). If learners prioritize, i.e., attach a particular value to learning an item or completing a learning task, the correlation between effort and monitoring judgments becomes positive. In

that case, *goal-driven* self-regulation takes place where "monitoring affects control" (i.e., MC model). In goal-driven self-regulation, learners allocate effort based on the importance of the items or their interest in it. In prior research, goal-driven self-regulation has been manipulated by increasing the relative importance of items, such as the number of points that can be obtained when the item is remembered correctly or by inducing time pressure (e.g., Ackerman 2014; Koriat et al. 2006). Goal-driven self-regulation has also been manipulated by giving learners a sense of agency, e.g., by asking learners how much effort they chose to invest instead of asking how much effort studying the item required (Koriat 2018).

Research has shown that children and adult learners can be steered to self-regulation in either a data- or goal-driven way by using incentives (e.g., Koriat et al., 2014a). For example, incentives can be provided by assigning 1 or 5 points to the correct recall of items. With higher incentives, the correlation between effort and monitoring judgments was positive instead of negative, indicating goal-driven learning. Also, by instructing students to adopt a facial expression that creates a feeling of effort (i.e., contracting eyebrows toward the center of the forehead), monitoring judgments were found to be lower, indicating a negative relation between experienced effort and monitoring (i.e., data driven). Yet, when adding time pressure to this situation, students started to learn in a goal-driven way and decided to allocate their study time to the easier items to recall as many items as possible at the end (Koriat and Nussinson 2009).

Koriat et al. (2014a) demonstrated an age-related increase in the ability to respond to data- and goal-driven manipulations in a task. Children between 14 and 15 years and college students were able to use data- and goal-driven self-regulation in the same task, whereas children aged 10–12 years could not use them both on the same task. In the current meta-analysis, we will examine the moderating effect of data- and goal-driven manipulations (e.g., manipulation of incentives, time pressure) on the strength and direction of the correlation between monitoring judgments and effort. Furthermore, we will examine the role of age differences by examining age and school level (i.e., grades 1–6, 7–12, or higher education) as moderators.

The Type of Effort Measurement

Looking at the different studies in which both monitoring and effort were measured, it is clear that effort can be measured in several different ways. Some studies use an objective measure of invested effort. Examples of objective measures are the study time needed to encode the learning materials (e.g., Ackerman 2014; Koriat et al. 2009a; Koriat and Ma'ayan, 2005), the time the participants needed to answer (i.e., response latency; Ackerman and Koriat, 2011; Ackerman and Zalmanov, 2012), or the number of trials needed before perfect recall (e.g., Koriat et al. 2009b). Other studies have measured effort by asking for learners' subjective ratings of mental effort (e.g., Baars et al. 2013; Kostons et al., 2012). For example, participants are asked to rate their experienced mental effort during a learning or test task after the task has been completed (Paas 1992). Presumably, there are more ways to conceptualize effort and measure it (Paas et al., 2003).

As the conceptualization and measurement of effort differ across the studies included in the current study, one of the questions in the present study is whether the type of effort measurement affects the relation between effort and monitoring. Firstly, Koriat (2018) showed that the level of agency reflected in mental effort rating (i.e., choosing to invest effort vs. rating

the required effort) influenced the direction of the association. Furthermore, mental effort ratings of required/invested effort may be more strongly correlated with monitoring judgments than objective measures as both are self-reported by the learner. We, therefore, examine if the different types of effort measurements relate in the same direction and with similar strength with monitoring judgments.

The Types of Monitoring Judgments

There are many different types of metacognitive monitoring judgments (see Dunlosky and Metcalfe 2008; Schraw 2009). Although the association between monitoring and effort has been found in various studies, different types of monitoring judgments were measured. Schraw (2009) describes three main categories of metacognitive monitoring judgments. Prospective judgments that are made before the task that is judged (i.e., predictions), concurrent judgments that are made during the performance on the task that is judged, or retrospective judgments which are made after completing the task that is judged (i.e., post-dictions). Examples of prospective judgments are judgments of learning (JOLs), feeling-of-knowing judgments (FOKs), and ease-of-learning judgments (EOLs). JOLs are often measured by asking learners to indicate the likelihood of remembering materials they just studied on a future test. FOKs can be measured by asking learners to predict whether they would be able to recognize the currently unrecallable information on a future test. EOLs are often measured by asking learners to indicate how easy or difficult it will be to learn certain learning materials (Dunlosky and Metcalfe 2008).

Examples of concurrent monitoring judgments are online confidence judgments, ease of solution judgments, and online performance accuracy judgments during an ongoing task (Schraw 2009). These types of judgments are made immediately after learners answer an item or perform a criterion task and require learners to rate their confidence in their answer, ease of problem solution, or performance accuracy. An important characteristic of concurrent judgments is that they are made on an item-by-item basis instead of over a set of items, which is typical for retrospective judgments. It, therefore, indicates a person's ability to judge their performance while it occurs.

Finally, retrospective judgments are, for example, EOLs and performance accuracy judgments made after a set of items or a criterion task is completed (Schraw 2009). Retrospective judgments can occur on an item-by-item level and global level. Still, they are always made after all items of all aspects of the criterion task have been completed. In the current meta-analysis, we examine if the strength of the association between monitoring judgments and effort is affected by the type of monitoring judgment that is measured.

The Type of Task

Next to the type of monitoring judgments and the type of effort measurement, the type of task a participant has to study, perform, or solve could be an interesting moderator of the relationship between effort and monitoring. There is evidence for the idea that effort informs monitoring for various types of tasks. For example, associations between effort and monitoring judgments have been found for studying word pairs (e.g., Koriat et al. 2009b), learning to solve problems (e.g., Baars et al. 2013), and other types of materials such as medical diagnoses (e.g., Blissett

et al., 2018). However, from the literature on the accuracy of monitoring judgments and how to improve them, it has become clear there are differences between different types of tasks in the accuracy of monitoring judgments and the effectiveness of interventions to improve this aspect (e.g., Ackerman and Thompson 2017; Baars et al. 2014; Thiede, Griffin, Wiley, and Redford 2009). For example, the delayed JOL effect was found to be most robust when studying word pairs (Rhodes and Tauber, 2011) but was not found for studying expository text (e.g., Maki 1998) or learning to solve problems (e.g., Baars et al. 2018). Therefore, it is examined if the strength of the correlation between monitoring judgments and invested effort is sensitive to different task types. In the current meta-analysis, most tasks concerned problem-solving tasks or learning words, word pairs, or other paired associates.

The Present Study

The current study aimed to assess the association between effort and monitoring by students in the context of studies on learning and performance. We used a meta-analytic approach to synthesize the results obtained in previous studies and to provide insight into the strength and direction of the estimated effect in the population. Specifically, the meta-analysis addresses the following questions:

1. What is the relation between students' effort and monitoring judgments during performance, learning, or training?
2. How do school level (i.e., grade 1–6, grade 7–12, higher education) and age influence the effect sizes?
3. How do goal-driven manipulations influence the effect sizes (e.g., incentives, time pressure)?
4. How do different types of effort measurements influence the effect sizes?
5. How do different types of monitoring judgments influence the effect sizes?
6. How does the type of task affect the effect sizes?

In line with the cue utilization perspective, we expected to find a negative association between monitoring judgments and invested effort (Hypothesis 1). Because some studies have found a weaker correlation for younger learners, we expected that the association would be weaker for children in grades 1–6, when compared with learners in grades 7–12, or higher education (Hypothesis 2).

We further expect that goal-driven regulation manipulations can influence the direction and strength of the association. A significant negative association is expected when self-paced study takes place, whereas a positive or nonsignificant association is expected when time pressure is applied (Hypothesis 3a). Furthermore, a significant negative association is expected when all items are equally important (i.e., no incentive). In contrast, a positive or nonsignificant association is expected when different incentives/points are awarded to the recall of different items in a learning task (Hypothesis 3b).

Furthermore, we hypothesized that mental effort ratings that express a sense of self-agency (choice to invest effort) would result in a positive association. In contrast, other mental effort ratings and objective effort measures will show a negative association with monitoring judgments (Hypothesis 4a). Additionally, we expected a stronger negative association for subjective mental effort ratings that ask learners to rate invested or required effort than for

objectively logged measures (Hypothesis 4b). For the type of monitoring judgments used, no particular differences were expected between confidence judgments, JOLs, or other metacognitive judgments. Furthermore, no specific differences were expected for the task type.

Method

Literature Search and Eligibility Criteria

A search was conducted in the internet databases ERIC (ProQuest interface) and Web of Science Core collection to locate relevant studies. We chose a time frame from 2000 to May 2020, because one of the first and often cited articles on the relation between effort and monitoring was published in 2008 (Koriat 2008). We conducted an initial search with the search terms "effort," "monitoring," and "learning." that was further restricted by only including English articles published in peer-reviewed journals in the field of education, educational, and cognitive psychology research. This first search resulted in 224 articles from Web of Science (WOS) and 146 articles from ERIC. A more expanded search was conducted on May 12, 2020 by including additional search terms for effort (i.e., effort* OR "response latency" OR "response time*" OR "study time"), monitoring (monitoring OR "judgment* of learning" OR "confidence judgment*" OR "confidence rating*" OR "metacognit* judgment*" OR "latency-confidence"), and learning (learning OR "self-regulation" OR "metacognition" OR "accuracy"). Again, search results were restricted by only including English, peer-reviewed articles. In WOS, the search was further restricted to only publications in psychological or educational sciences research domains. The second search resulted in 384 articles in WOS and 211 articles in ERIC. After removing duplicates from the initial and expanded search, 675 articles remained. Furthermore, we checked the references of selected studies for additional studies.

To select all relevant studies on the association between effort and monitoring, specific criteria for inclusion were developed.

1. The (cor) relation between effort and monitoring judgments and the sample size was reported or received after a request via email to the corresponding author.
2. Both effort and monitoring judgments were measured on a quantitative scale in the context of learning or performance.
3. Effort and monitoring were measured in one study or experiment in the same trial for the same item or criterion task. If there were multiple parts of the study in which effort and monitoring were measured (e.g., pretest, learning phase, and posttest), data from the learning phase and/or posttest were used.
4. The description of the measurement of both effort and monitoring judgments is reported in such a way that it could be coded what type of effort measurement and monitoring judgments were used in the study.

As can be seen in Fig. 1, from the 675 articles found, 617 articles were immediately excluded based on criteria 2–4, whereas 58 articles were selected for further coding. Furthermore, based on a snowball search (e.g., references in selected articles), an additional 16 articles were identified. Of the 74 articles selected for further coding, five were removed because they did not meet criteria 2–4 (e.g., effort and monitoring judgment not measured in the same trial).

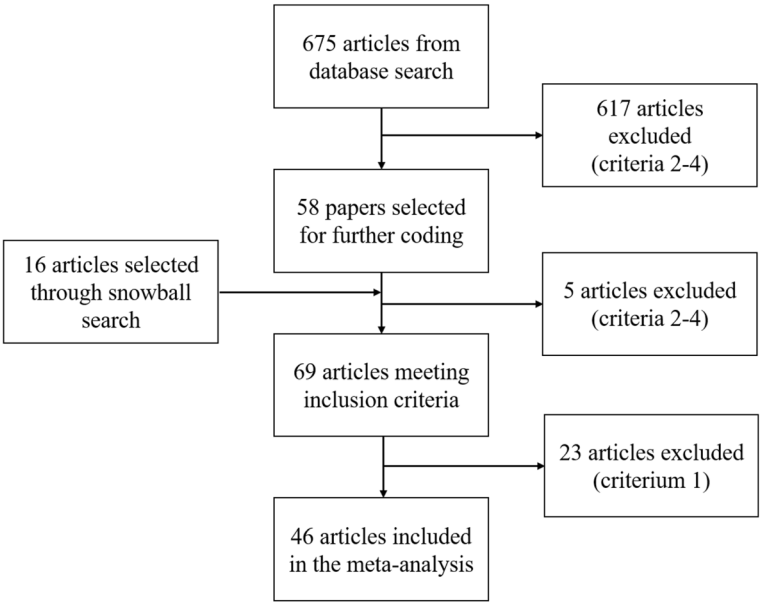

Fig. 1 Flow chart of selection of articles

Additionally, 23 articles were excluded because the correlation between monitoring judgment and effort could not be obtained. The search and selection process resulted in a final subset of 46 articles with 164 correlations between effort and monitoring found in 123 participant groups with a total of 5819 participants (see Table 3 Appendix).

Coding

A coding scheme was used to describe the articles included in the current study. In addition to effect size data (i.e., r and N), several moderators were coded.

Sample Characteristics

Because research suggests both data- and goal-driven self-regulation have been sensitive to age-differences, we coded the school level (i.e., grades 1–6, grades 7–12, or higher education) of the sample. Also, the mean age of the participant sample was coded. For some higher education samples, the mean age was not reported. For these studies, we used the average age of the other included higher education samples ($M = 23.02$).

Data- and Goal-Driven Manipulations

Data- and goal-driven self-regulation can co-occur in the same task (Koriat et al., 2014a). Therefore, the presence of goal-driven manipulations was coded. Goal-driven self-regulation is often elicited by using time pressure or item incentives, but can also be manipulated by promoting a sense of self-agency by asking learners how much effort they chose to invest (Koriat 2018). If participants experienced one or more of these three elements, it was coded that a goal-driven manipulation was present. We further included separate variables on the

presence of time pressure during the learning or performance phase (compared with sufficient or self-paced study time) and the presence of incentives (i.e., differential point distribution for correctly recalling an item or solving a task) or not.

Effort Measurement

We further made a distinction between effort measures that were (a) objectively logged (i.e., time or number of trials), (b) subjective ratings of invested/required mental effort, or (c) subjective ratings of learners' choice to invest effort (i.e., measures to promote self-agency and goal-driven self-regulation; Koriat 2018). Objectively logged measures were further subdivided in study time, response latency, and the number of trials needed before perfect recall/acquisition.

Monitoring Judgments

We coded type of monitoring judgment (i.e., JOLs, confidence ratings, or other measures) and the timing of the judgment (i.e., prospective, concurrent, retrospective) according to the descriptions we have provided earlier in the paper.

Type of Task

We coded task type with the categories: (a) word learning and paired associations, (b) problem solving, and (3) other. The first category included studies on learning Chinese words (e.g., Jia et al. 2016), word pairs (e.g., Koriat et al. 2006), or other paired associates (e.g., Ackerman and Koriat, 2011). The category problem solving included, for example, hereditary problem solving in biology (e.g., Baars et al. 2014), misleading math/reasoning problems (e.g., Ackerman and Zalmanov 2012), and compound remote associates problems (e.g., Ackerman 2014). The "other" category included tasks such as diagnosing medical cases (Blissett et al. 2018), reading/studying a text (e.g., Kostons and De Koning 2017), or a general knowledge test (Koriat and Ackerman 2010a).

Data Analyses

All analyses were performed in Comprehensive Meta-Analysis statistical software (version 3.0.1.0; Biostat, Englewood, NJ; Borenstein et al., 2009). The correlation was used as the effect size measure based on the correlation and sample size reported in the articles or retrieved from the authors of the study. Most studies concerned experiments that reported the correlation between effort and monitoring judgment per experimental condition; of some studies, only the correlation for the whole study was available. If a study reported several correlations of the same participants (i.e., two or more correlations of participants from the same condition), a combined, mean effect size was computed. The mean correlation was estimated using a random-effects model. To assess statistical heterogeneity, we calculated the Q and I^2 statistics (Borenstein et al. 2009; Higgins and Thompson 2002). The I^2 is an index of heterogeneity in percentages (i.e., 25% = low, 50% = moderate, 75% = high heterogeneity). Moderator analyses for the categorical variables were conducted based on analyses of variances (ANOVAs). Between-group differences in the categorical mixed-effects analyses were tested with the Q statistic for between group means. Furthermore, we conducted a random effects meta-

regression model (using method of moments) to examine the effect of age (see Borenstein et al. 2009). Finally, we conducted a random effects meta-regression model (using method of moments) in which multiple moderators were included to test which moderators remained significant after controlling for the effects of other moderators. Additionally, we assessed publication bias.

Results

The effect size reported is a correlation coefficient (r), for which values of .10 are considered small, .30 medium, and .50 large effects (Cohen 1988).

Research Question 1: The Relation Between Monitoring and Effort

Because in some studies more than one correlation was reported for the same group of participants, a combined effect size was calculated in which the mean of the outcomes is used for the analysis. To answer Research Question 1, we analyzed the mean correlation between effort and monitoring judgment ($k = 123$). In support of Hypothesis 1, a negative, small correlation was found, $r = -.355$ (95% CI [−.408, −.300]). The effect was heterogenous, $Q(122) = 597.12$, $p < .001$, $I^2 = 79.57$, $T^2 = 0.09$ ($SE = .02$).

Research Questions 2–6: Moderators

Table 1 presents the results of the moderator analyses. To answer Research Question 2, we examined the effect of age and school level on the association between effort and monitoring judgments. In contrast to Hypothesis 2, results from the meta-regression revealed that the mean age of the participant sample was not a significant predictor of the correlation between effort and monitoring judgments, $b = .009$ (SE = .005), $Q(1) = 2.95$, $p = .086$. However, results from the moderator analysis with school level as a categorical variable showed that school level was a significant moderator of the relationship between effort and monitoring, $Q(2) = 14.66$, $p = .001$. The results demonstrated a higher negative correlation between effort and monitoring judgments for grades 7–12 when compared with grades 1–6 and higher education. These results suggest there is no linear effect of school level on the association between effort and monitoring.

To answer Research Question 3, we examined the effect of goal-driven manipulation on the correlations between effort and monitoring judgments. Firstly, we compared the overall effect of the presence of goal-driven manipulators on the effect size, such as the presence of incentives, time pressure, or self-agency manipulations. The analysis revealed a significant moderation effect, $Q(1) = 6.39$, $p = .011$. Although there was still a significant negative correlation between effort and monitoring judgments, the correlation was lower than when goal-driven self-regulation was not manipulated. We further examined the effect of using incentives and time pressure separately. In support of Hypotheses 3a and 3b, we found that the association between effort and monitoring judgments became nonsignificant when incentives or time pressure were used (see Table 1).

We further examined the effects of mental effort measures. As mentioned, goal-driven self-regulation can be manipulated by the question that is asked when participants rate their effort by promoting self-agency (i.e., how much effort did you choose to invest?). We examined the effect of self-agent mental effort ratings compared with objectively logged effort ratings, and other subjective mental effort ratings. Because the study by Koriat et al. (2014b) included two

Table 1 Results of the moderator analyses (mixed-effects model)

Moderator	*k*	*r*	95% CI	*z*	*Q*	df	*p*
School level					14.66	2	.001
Grades 1–6	32	− .349	[− .447, − .243]	− 6.14***			
Grades 7–12	18	− .553	[− .645, − .445]	− 8.47***			
Higher education	73	− .296	[− .363, − .226]	− 7.95***			
Goal-driven manipulation					3.39	1	.011
Not present	102	− 0.385	[− .440, − .327]	− 12.01***			
Present	21	− 0.193	[− .332, − .046]	− 2.57*			
Time pressure					5.55	1	.019
Time pressure	10	− .117	[− .325, .102]	− 1.05			
Self-paced/sufficient	113	− .373	[− .426, −.318]	− 12.02***			
Incentives					4.34	1	.037
No incentives	111	− .374	[− .428, − .354]	− 12.02***			
Incentives	12	− .174	[− .354, .018]	−1.76			
Effort measure					25.38	2	< .001
Objective measure	89	− .310	[− .371, − .247]	− 9.10***			
Self-agent rating	3	.283	[− .078, .578]	1.55			
Other ME rating	31	−.502	[− .578, − .418]	− 10.09***			
Effort measure					42.50	4	< .001
Study time	54	− .210	[− .291, − .126]	− 4.84***			
Response latency	30	− .464	[− .548, − 369]	− 8.61***			
Number of trials	5	− .337	[− .571, − .052]	− 2.30*			
Self-agent rating	3	.284	[− .066, .571]	1.60			
Other ME rating	31	− .501	[− .575, −.420]	− 10.41***			
Monitoring judgment					14.12	2	.001
Confidence rating	30	− .488	[− .573, − .393]	− 8.86***			
JOL	82	347	[− .432, − .216]	− 8.00***			
Other	11	− .457	[− .581, − .312]	− 5.67***			
Timing monitoring judgment					16.71	1	< .001
Concurrent	39	− .492	[− .562, −.562]	− 10.94***			
Prospective	84	− .281	[− .343, −.217]	− 8.25***			
Type of task					32.20	2	< .001
Problem solving	38	− .533	[− .597, − .462]	− 12.33***			
Word/paired associated	71	− .250	[− .318, − .179]	− 6.76***			
Other task	14	− .267	[− .408, − .113]	− 3.35**			

ME mental effort

*$p < .05$

**$p < .01$

***$p < .001$

types of effort measures (i.e., objective self-study time measure and self-agent/other mental effort ratings) for the same participants, we either had to exclude the study from the analysis or only use the data from one of the measures to be able to include it in the moderator analysis. Because the role of self-agent mental effort ratings was only examined in a few studies, we excluded the correlations resulting from the objective measure of this study to ensure that the sample did not appear twice in the analysis. In support of Hypothesis 4a, mental effort ratings that promote a sense of self-agency result in a nonsignificant positive association between effort and monitoring judgments, in line with goal-driven self-regulation. In support of Hypothesis 4b, subjective mental effort ratings resulted in a higher relationship relative to objectively logged effort ratings. When we made a further distinction between the type of objective effort measure, results revealed that study time measures resulted in a lower correlation than response latency measures (see Table 1).

To answer Research Question 5, we examined the effect of the type of monitoring judgment that was used. Type of monitoring judgment (i.e., JOL, confidence rating, other) was a significant moderator to the relation between effort and monitoring (see Table 1). Although a significant, negative correlation was found for all judgment types, the association was smaller for JOLs. We further made a distinction between the timing of the judgment and found that concurrent judgments resulted in a higher negative correlation than prospective judgments.

Finally, we examined the role of task type (i.e., problem solving, word learning/paired associates, and other). Because the study by Dentakos, Saoud, Ackerman, and Toplak (2019) included multiple task types, we only included the two problem-solving tasks. We excluded the other task (i.e., answering general knowledge questions from the analysis). It seemed that problem-solving tasks had a stronger negative correlation compared with the other types of tasks.

Meta-regression

For several moderators, significant effects were found. However, closer inspection of Table 3 Appendix reveals that some moderators share substantial overlap. For example, most of the studies conducted in secondary education (i.e., grades 7–12) used problem-solving task. Therefore, we conducted a meta-regression with multiple moderators (i.e., one moderator per research question), to examine which of the moderators had a unique effect on the correlation between effort and monitoring judgments controlling for the effect of other moderators. Table 2 presents the results of the meta-regression. Again, from the study by Koriat et al. (2014b), only the mental effort ratings were included, and from Dentakos et al. (2019), only the problem-solving tasks.

The model including all moderators (excluding the intercept) was significant, $Q(7) = 51.86$, $p < .001$, $R^2 = .39$. The goodness of fit test showed that the covariates in the model did not explain all heterogeneity, $T^2 = .05$, $Q(115) = 380.68$, $p < .001$. School level and mental effort measure (i.e., subjective vs. objective measure) were no longer significant predictors when other moderators were included in the model. Goal-driven manipulations had a significant effect. When a study manipulated goal-driven self-regulation, the correlation between effort and monitoring judgments became less negative. Also, prospective monitoring judgments resulted in a weaker negative correlation relative to concurrent judgments. Finally, problem-solving tasks resulted in a stronger negative correlation compared with other tasks.

Table 2 Meta-regression

	b	*SE*	95% CI	*z*	*p*	*Q*
Intercept	− .357	.09	[− .536, − .178]	− 3.91	< .001	
School level						2.84
Grades 7–12 vs. grades 1–6	.017	.12	[− .211, .246]	0.15	.882	
HE vs. grades 1–6	.100	.07	[− .033, .234]	1.47	.142	
Goal manipulation vs. no manipulation	.190	.07	[.044, .335]	2.55	.011	
Subjective vs. objective effort	− .072	.08	[− .236, .093]	− 0.85	.395	
Prospective vs. concurrent judgments	.171	.08	[.023, .312]	2.27	.023	
Task						11.15**
Problem solving vs. other	− .340	.10	[− .544, − .136]	− 3.27	.001	
Paired associates vs. other	− .154	.11	[− .361, .034]	− 1.45	.147	

For moderators with more than two categories, the combined effect is tested with the *Q* statistic

**$p < .01$

Publication Bias

To assess publication bias, we inspected the funnel plot by plotting each individual study effect size against its standard error, Egger's regression intercept (Egger et al., 1997), and Duval and Tweedie's (2000) trim-and-fill technique, and conducted a classic fail-safe *N* analysis. The fail-safe *N* estimated the number of studies with an effect size of zero that are required to nullify the overall effect size. See Fig. 2 for the funnel plot. Egger's linear regression test for asymmetry did not suggest publication bias, $t(121) = 0.39$, $p = .348$. Duval and Tweedie's trim-and-fill technique (31 studies trimmed at the right side) resulted in an adjusted correlation from − .355 to − .227 (95% CI [− .291, − .162]. The fail-safe *N* suggested that 20,731 missing studies are needed for the result of this meta-analysis to be nonsignificant ($p > .05$).

Discussion

Research has shown that without any additional instructional support, learners experience difficulties in making accurate monitoring judgments (e.g., Ackerman and Thompson 2017; Baars et al. 2014; Thiede et al. 2009). As a result of this, students' regulation of further learning is harmed, and thereby learning outcomes decreased (e.g., Dunlosky and Rawson 2012). Hence, it is crucial to know more about how students make monitoring judgments and specifically what cues they use as a basis for their monitoring judgments to support effective monitoring and regulation during self-regulated learning. In the current study, the association between effort and monitoring judgments made by students in the context of learning and performance was investigated. Furthermore, the role of possible moderators was examined. Using a meta-analytic approach, we integrated the results from previous studies on effort and monitoring to provide insight into the strength and direction of the estimated effect in the population.

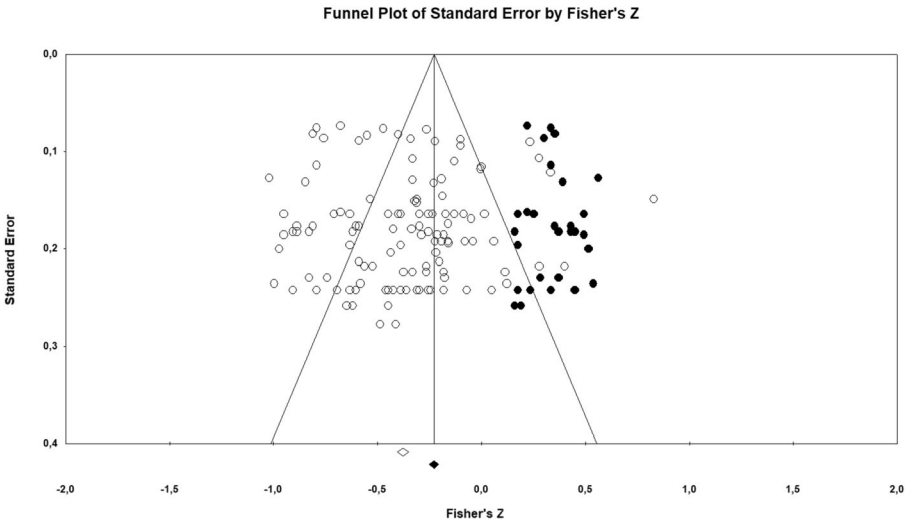

Fig. 2 Funnel plot with observed and imputed studies. The white dots represent the observed study samples included in the meta-analysis ($k = 123$), and the black dots represent the 31 studies trimmed at the right side using Duval and Tweedie's trim-and-fill technique, resulting in an adjusted correlation from − .355 to − .227 between effort and monitoring judgments

Main Findings

The results showed a negative, medium correlation between effort and monitoring judgments ($r = -.355$). These results are in line with the cue utilization perspective (Koriat 1997), in which effort is described as a potential cue. That is, the perceived invested effort in a learning task is supposedly used as a cue to make monitoring judgments. Furthermore, several moderators were examined. The role of age and school level was investigated because earlier studies demonstrated age-related improvements in cue utilization (e.g., Hoffmann-Biencourt et al. 2010; Koriat et al. 2009a). In our meta-analysis, we did not find age-related differences in the correlation between effort and monitoring judgments. Koriat et al. (2009a) showed that the critical development in the reliance of the memorizing effort heuristic develops somewhere in the third grade. In our study, we only were able to include a few young samples in our analyses (e.g., learners in grades 1 and 2); this might explain why no age-related differences were found.

Concerning data- and goal-driven self-regulation, the meta-analysis provides evidence for both types of processes. Overall, learners tend to focus on data-driven self-regulation in which monitoring judgments are based on the amount of effort that was needed to learn the study material or to solve the problem, as indicated by the negative correlations between effort and monitoring judgments. However, the results of the moderator analyses and meta-regression showed that the use of incentives, time pressure, or promoting feelings of self-agency resulted in a nonsignificant correlation between effort and monitoring judgments. These results suggest that students use data- as well as goal-driven self-regulation (e.g., Koriat et al., 2014a; Koriat and Nussinson 2009). However, a significant positive correlation was not obtained in the moderator analyses, which suggests that it is challenging to promote goal-driven self-regulation in students.

We furthermore examined the role of differences in the measurement of effort and monitoring judgments. We hypothesized that mental effort ratings of invested/required effort would result in a stronger negative association with monitoring judgments than when the effort was objectively logged (e.g., study time and response latency), because effort ratings and monitoring judgments were both self-reported by the learner. In our initial moderator analysis, we found evidence for this hypothesis, but the effect of effort measures disappeared when other moderators were included in the analysis (see Table 2).

Type of monitoring judgment (i.e., JOL, confidence rating, other) was found to be a significant moderator to the relation between effort and monitoring. That is, the correlation was weaker for prospective JOLs compared with concurrent confidence judgments and other judgments. When we examined the effect of timing, we found evidence that prospective judgments resulted in a weaker correlation compared with concurrent judgments. We did not have prior expectations about differences in monitoring judgments. Possibly, the phrasing or the timing (i.e., concurrent vs. prospective) of monitoring judgments prompts learners to use effort as a cue to a certain extent. For example, concurrent judgments often ask learners to rate their confidence in their answers or to self-assess how well they have performed a certain task; these judgments are typically measured during a performance/test phase. In contrast, prospective judgments are more focused on future recall or performance and are measured during the learning phase. Possibly, learners rely more on effort during performance when compared with learning phases.

Interestingly, the type of task (i.e., problem solving, word learning/paired associates, other tasks) was found to be a significant moderator of the relationship between effort and monitoring judgments. Specifically, results showed that the negative correlation was higher for problem-solving tasks relative to learning words or paired associates and other tasks (e.g., reading). We did not have prior expectations about the effect of different tasks. Perhaps

specific processes or features of the task affect the use of effort as a cue. Possibly learners believe that effort is a better cue for judging how well you (will) perform on a problem-solving task than for how well you can recall words or paired associated in the future.

Limitations and Future Studies

The current study has some limitations that should be taken into consideration when interpreting the findings. One limitation is that we did not include "gray literature." Future research on effort and monitoring using review and meta-analysis could benefit from a more in-depth search also covering dissertations, conference papers, or other reports. Furthermore, the role of the moderators that were tested in the current study requires more attention. Firstly, it remains unclear why certain types of tasks, such as problem solving, yield a higher negative correlation compared with others. Furthermore, it is unclear why concurrent judgments result in a stronger negative correlation than prospective judgments. Future studies could examine this further by examining the association between effort and monitoring judgments in a within-subjects design in which effort and monitoring judgments are measured for different task types and during different phases (i.e., learning phases and performance phases).

Concerning the school level, earlier work has shown that primary education students showed a smaller correlation indicating a developmental trajectory in using effort as a cue for monitoring (Koriat et al. 2009b). More research with younger learners (e.g., learners in grade 1) will give more insight into age-related differences in cue utilization. Although our meta-analysis revealed evidence for both goal- and data-driven self-regulation, in our study, we were only able to include a small number of studies in which goal-driven self-regulation was manipulated. With more future studies on goal vs. data-driven scenarios during learning, future meta-analyses could further investigate the moderating role of goal- vs. data-driven self-regulation in the correlation between effort and monitoring.

Furthermore, although many studies have shown a negative linear correlation between effort and monitoring, some studies reported an inverted U-shaped curvilinear relationship between effort and monitoring, such as between study time and JOLs (see Undorf and Ackerman 2017). This curvilinear relationship could not be explained by a data-driven or goal-driven approach alone. In their study, Undorf and Ackerman (2017) investigated different models for study time allocation (i.e., Discrepancy reduction model, DRM, Nelson and Narens 1990; Region of proximal learning model, RPL, Metcalfe and Kornell 2005; Diminishing criterion model, DCM, Ackerman 2014) to explain the curvilinear findings. The results showed that learners set time for learning an item (i.e., a criterion), and after this time had passed, the relationship between study time and monitoring judgments changed. These results confirmed the DCM model (Ackerman 2014) which predicts that for more complex learning tasks, such as problem-solving tasks, learners invest effort in a goal-driven way at first but after time passes the goal could be compromised, and the relation between effort and monitoring becomes negative (i.e., data driven). These results suggest there is a different type of relation between effort and monitoring compared with the relation found in the current meta-analysis. Future studies could investigate this curvilinear relationship between effort and monitoring and advance our understanding of effort as a cue using multilevel modeling techniques.

The main finding of this meta-analysis is a negative correlation between effort and monitoring, which suggests effort is being used as a cue to make monitoring judgments. However, we did not investigate whether effort is a good cue for performance (i.e., cue

diagnosticity); neither did we examine monitoring accuracy. For example, Raaijmakers et al. (2017) found that feedback valence alters mental effort ratings. This could mean that the invested effort is not a good predictor of performance. Yet, because monitoring judgments are inferential, their accuracy depends on the relation between the cue and performance (Koriat 1997). In a future study, meta-analytic structural equation modeling could be conducted in which cue diagnosticity, cue utilization, and monitoring accuracy are investigated in the same analysis (see Dunlosky et al., 2016).[1] Furthermore, according to cognitive load theory (CLT; Sweller et al., 1998, 2019), two main types of cognitive load are affecting the learning processes differently, i.e., intrinsic and extraneous cognitive load. Intrinsic load caused by the learning material itself is inherent to the material and the learning process. If perceived effort would be based on this type of cognitive load, it could potentially be a valid cue for monitoring and self-regulated learning as a whole. That is, if the effort is too high or too low, learning is probably not optimal. Extraneous load is caused by the design of the learning materials, which does not aid the learning process. If this type of load contributes to perceived effort, it could blur the relationship between effort and learning because it increases invested effort without adding to learning performance. This would leave the learner with a very complicated situation of perceiving effort and using that as a cue to their self-regulated learning process. Future research could look into how different types of cognitive load affect perceived effort and if they are being used as a cue for monitoring.

Conclusion

The current study was the first to investigate the association between effort and monitoring using a meta-analytic approach. The findings showed that there is a medium, negative correlation between effort and monitoring judgments suggesting effort is used as a cue for monitoring. Interestingly, the type of monitoring judgment (i.e., concurrent confidence ratings vs. prospective JOLs), the type of task, and goal-driven manipulations (e.g., incentives, time pressure) moderate this relation. This can have important implications for future research on the use of effort as a cue for monitoring in self-regulated learning.

Acknowledgments We would like to thank Corien Woudenberg and Jonna Kirveskoski for their help with the data analysis.

Authors' Contributions Martine Baars and Lisette Wijnia equally contributed to the manuscript and therefore share first authorship.

Compliance with Ethical Standards

Conflicts of Interest/Competing Interests The authors declare that they have no conflict of interest..

Availability of Data and Material Not applicable.

Code Availability Not applicable.

[1] We examined if absolute monitoring accuracy score (0–100%) moderated the association between effort and monitoring in a sample of 26 studies. The results showed that monitoring accuracy score did not affect the relationship between effort and monitoring judgments, $b = .007$, $z = 0.60$, $p = .551$.

Appendix

Table 3 Summary of coded studies and associated effect sizes

Article	Sample	Level	Goal driven	Monitoring (timing)	Effort	Task	*r*	*n*
1. Ackerman (2014)	Exp. 1	HE	–	JOL (p)	ST	Paired associates	− .21	32
	Exp. 2-incentive	HE	Incentive	CR (c)	RL	Problem solving	− .61	40
	Exp. 2-no incentive	HE	–	CR (c)	RL	Problem solving	− .72	20
	Exp. 3-ample time	HE	–	CR (c)	RL	Problem solving	− .68	22
	Exp. 3-time pressure	HE	Pressure	CR (c)	RL	Problem solving	− .60	20
	Exp. 4-ample time	HE	–	CR (c)	RL	Problem solving	− .63	22
	Exp. 4-time pressure	HE	Pressure	CR (c)	RL	Problem solving	− .53	25
	Exp. 5-forced-report	HE	Pressure	CR (c)	RL	Problem solving	− .48	24
	Exp. 5-free-report	HE	Pressure	CR (c)	RL	Problem solving	− .41	27
2. Ackerman and Koriat (2011)	Exp. 1-2nd grade	1–6	–	CR (c)	RL	Paired associates	− .57	18
	Exp. 1-5th grade	1–6	–	CR (c)	RL	Paired associates	− .55	18
	Exp. 2-2nd grade	1–6	–	CR (c)	RL	Paired associates	− .54	20
	Exp. 2-5th grade	1–6	–	CR (c)	RL	Paired associates	− .66	20
	Exp. 3-2nd grade	1–6	–	CR (c)	RL	Other: questions about narrated slideshow	− .42	40
	Exp. 3-5th grade	1–6	–	CR (c)	RL	Other (questions about narrated slideshow)	− .56	40
3. Ackerman and Zalmanov (2012)	Exp. 1-MC test	HE	–	CR (c)	RL	Problem solving	− .54	35
	Exp. 1-Open-ended	HE	–	CR (c)	RL	Problem solving	− .40	34
	Exp. 2	HE	–	CR (c)	RL	Problem solving	− .75	28
4. Baars et al. (2018)	Delayed JOL	1–6	–	JOL (p)	ME	Problem solving	− .59	41
	Immediate JOL	1–6	–	JOL (p)	ME	Problem solving	− .67	35
5. Baars et al. (2014)	Exp. 1-training	7–12	–	JOL (p)	ME	Problem solving	− .26	23
	Exp. 1-control	7–12	–	JOL (p)	ME	Problem solving	− .76	21
	Exp. 2-control	7–12	–	JOL (p)	ME	Problem solving	− .71	35
	Exp. 2-training only	7–12	–	JOL (p)	ME	Problem solving	− .74	32
	Exp. 2-standards only	7–12	–	JOL (p)	ME	Problem solving	− .55	33
	Exp. 2-standards + training	7–12	–	JOL (p)	ME	Problem solving	− .71	33
6.	Completion problems	7–12	–	JOL (p)	ME	Problem solving	− .68	33

Table 3 (continued)

Article	Sample	Level	Goal driven	Monitoring (timing)	Effort	Task	*r*	*n*
Baars et al. (2013)								
	Worked-out examples	7–12	–	JOL (p)	ME	Problem solving	− .72	33
7. Baars and Wijnia (2018)	Study 1	7–12	–	Other (c)	ME	Problem solving	− .66	178
	Study 2	7–12	–	Other (c)	ME	Problem solving	− .50	147
8. Baars et al. (2017)	One study	7–12	–	Other (c)	ME	Problem solving	− .53	130
9. Ball et al. (2014)	Exp. 1a	HE	–	JOL (p)	ST	Paired associates	− .53	35
	Exp. 1b	HE	–	JOL (p)	ST	Paired associates	− .56	29
	Exp. 2	HE	–	JOL (p)	ST	Paired associates	− .37	29
	Exp. 3	HE	–	JOL (p)	ST	Paired associates	− .29	35
10. Blissett et al. (2018)	One study	HE	–	CR (c)	ME	Other (diagnosing medical cases)	− .18	22
11. Burkett and Azevedo (2012)	Combined	HE	–	JOL (p)	ST	Other (studying science with multimedia)	.02	40
12. Cuevas et al. (2002)	One study	HE	–	JOL (p)	ME	Other (studying/training tutorial)	− .69	61
13. Dentakos et al. (2019)	Combined	HE	–	CR (p)	ME	Problem solving and other (general knowledge)	− .33	136
14. Heijltjes et al. (2015)	Combined	HE	–	CR (c)	ME	Problem solving	− .67	152
15. Hoffmann-Biencourt et al. (2010)	Group 1	1–6	–	JOL (p)	ST	Paired associates	− .13	40
	Group 2	1–6	–	JOL (p)	ST	Paired associates	− .17	40
	Group 3	1–6	–	JOL (p)	ST	Paired associates	− .29	40
	Group 4	7–12	–	JOL (p)	ST	Paired associates	− .38	40
16. Hoogerheide et al. (2014)	Exp. 1	7–12	–	JOL (p)	ME	Problem solving	.00	78
	Exp. 2	7–12	–	JOL (p)	ME	Problem solving	− .10	134
17. Jia et al. (2016)	Exp. 2a	HE	–	JOL (p)	ST	Word learning	− .04	30
18. Koriat (2008)	Exp. 4	HE	–	JOL (p)	Trials	Paired associates	− .25	20
19. Koriat (2018)	Exp. 1-data driven	HE	Incentive + Pressure	JOL (p)	ME	Paired associates	−.49	48
	Exp. 1-goal-driven	HE	Incentive, Pressure + Agency	JOL (p)	ME-A	Paired associates	.68	48
	Exp. 2-data driven	HE	–	JOL (p)	ME	Paired associates	− .74	40
	Exp. 2-goal-driven	HE	Agency	JOL (p)	ME-A	Paired associates	− .25	40
20. Koriat and Ackerman (2010a)	Exp. 1-2nd grade	1–6	–	CR (c)	RL	Other (general knowledge)	− .30	20
	Exp. 1-3rd grade	1–6	–	CR (c)	RL	Other (general knowledge)	− .43	20

Table 3 (continued)

Article	Sample	Level	Goal driven	Monitoring (timing)	Effort	Task	*r*	*n*
	Exp. 1-5th grade	1–6	–	CR (c)	RL	Other (general knowledge)	− .54	20
	Exp. 2-2nd grade	1–6	–	CT (c)	RL	Other (general knowledge)	− .26	24
	Exp. 2-5th grade	1–6	–	CR (c)	RL	Other (general knowledge)	− .51	24
21. Koriat and Ackerman (2010b)	Exp. 1	HE	–	JOL (p)	ST	Paired associates	− .40	20
	Exp. 2	HE	–	JOL (p)	ST	Paired associates	− .20	25
22. Koriat et al. (2014a)	Exp. 1-combined	1–6	Incentive	JOL (p)	ST	Paired associates	− .09	40
	Exp. 2-combined	1–6	Incentive	JOL (p)	ST	Paired associates	− .23	40
	Exp. 3-combined	1–6	Incentive	JOL (p)	ST	Paired associates	− .16	40
	Exp. 4-combined	7–12	Incentive	JOL (p)	ST	Paired associates	− .35	20
	Exp. 5-combined	1–6	Incentive	JOL (p)	ST	Paired associates	− .23	60
23. Koriat et al. (2009a)	1st grade	1–6	–	JOL (p)	ST	Paired associates	− .07	20
	2nd grade	1–6	–	JOL (p)	ST	Paired associates	.05	20
	3rd grade	1–6	–	JOL (p)	ST	Paired associates	− .24	20
	5th grade	1–6	–	JOL (p)	ST	Paired associates	− .29	20
	6th grade	1–6	–	JOL (p)	ST	Paired associates	− .18	20
24. Koriat et al. (2009b)	Exp. 1-2nd grade	1–6	–	JOL (p)	Trials	Paired associates	− .56	20
	Exp. 1-4th grade	1–6	–	JOL (p)	Trials	Paired associates	− .60	20
	Exp. 2-2nd grade	1–6	–	JOL (p)	Trials	Paired associates	− .04	30
	Exp. 2-4th grade	1–6	–	JOL (p)	Trials	Paired associates	− .22	30
25. Koriat and Ma'ayan (2005)	Exp. 1	HE	–	JOL (p)	ST	Paired associates	− .22	27
	Exp. 2	HE	Pressure	JOL (p)	ST	Paired associates	.11	23
26. Koriat et al. (2006)	Exp. 1	HE	–	JOL (p)	ST	Paired associates	− .42	20
	Exp. 3-combined	HE	–	JOL (p)	ST	Paired associates	− .32	34
	Exp. 4	HE	–	JOL (p)	ST	Paired associates	− .60	20
	Exp. 5-control	HE	–	JOL (p)	ST	Paired associates	− .39	16
	Exp. 5-incentive (combined)	HE	Incentive	JOL (p)	ST	Paired associates	− .45	16
	Exp. 6-control	HE	Pressure	JOL (p)	ST	Paired associates	.38	24
	Exp. 6-incentive (combined)	HE	Pressure + Incentive	JOL (p)	ST	Paired associates	.27	24
	Exp. 7-combined	HE	Incentive	CR (c)	RL	Problem solving	− .30	46
27. Koriat et al. (2014b)	Data driven (combined)	HE	–	JOL (p)	ST & ME	Paired associates	− .52	21
	Goal-driven (combined)	HE	Agency	JOL (p)	ST & ME-A	Paired associates	− .12	21

Table 3 (continued)

Article	Sample	Level	Goal driven	Monitoring (timing)	Effort	Task	*r*	*n*
28. Kostons and De Koning (2017)	One study	1–6	–	JOL (p)	ME	Other (reading)	− .10	116
29. Kostons et al. (2012)	One study	7–12	–	Other (c)	ME	Problem solving	− .66	80
30. Lachner et al. (2020)	Exp. 1	HE	–	JOL (p)	ME	Other (studying)	.27	91
	Exp. 2	HE	–	JOL (p)	ME	Other (studying)	.23	126
31. Li et al. (2016)	Exp. 2a	HE	–	JOL (p)	ST	Word learning	− .05	38
32. Miele et al. (2011)	Exp. 1-combined	HE	–	JOL (p)	ST	Paired associates	− .01	75
33. Mihalca et al. (2017)	One study	HE	–	Other (p)	ME	Problem solving	− .13	86
34. Mueller et al. (2014)	Exp. 2-self-paced	HE	–	JOL (p)	ST	Paired associates	.06	30
35. Mueller et al. (2016)	Exp. 1-self-paced	HE	–	JOL (p)	ST	Paired associates	− .08	30
	Exp. 2-self-paced	HE	–	JOL (p)	ST	Paired associates	− .16	36
36. Paik and Schraw (2013)	One study	HE	–	JOL (p)	ME	Problem solving	− .77	65
37. Price and Harrison (2017)	Exp. 3-combined	HE	–	JOL (p)	ST	Paired associates	.32	71
38. Raaijmakers et al. (2019)	Study 1	7–12	–	Other (c)	ME	Problem solving	− .59	189
	Study 2	7–12	–	Other (c)	ME	Problem solving	− .64	137
39. Roebers et al. (2019)	2nd grade	1–6	–	CR (c)	RL	Paired associates	− .38	150
	4th grade	1–6	–	CR (c)	RL	Paired associates	− .44	175
40. Schnaubert and Bodemer (2017)	One study	HE	–	CR (c)	ME	Other (studying)	− .32	63
41. Thompson et al. (2011)	Exp. 1	HE	–	Other (c)	RL	Problem solving	− .32	90
	Exp. 2	HE	–	Other (c)	RL	Problem solving	− .30	48
	Exp. 3	HE	–	Other (c)	RL	Problem solving	− .22	128
	Exp. 4	HE	–	Other (c)	RL	Problem solving	− .19	64
42. Undorf and Ackerman (2017)	Exp. 1-combined	HE	Incentive	JOL (p)	ST	Paired associates	− .19	50
	Exp. 2-standard	HE	–	JOL (p)	ST	Paired associates	− .28	32
	Exp. 2-accuracy	HE	–	JOL (p)	ST	Paired associates	− .19	30
	Exp. 3-standard	HE	–	JOL (p)	ST	Paired associates	− .16	30
	Exp. 3-time pressure	HE	Pressure	JOL (p)	ST	Paired associates	− .22	30
43. Undorf and Erdfelder (2011)	Learning	HE	–	JOL (p)	ST	Paired associates	− .32	23
	Observing	HE	–	JOL (p)	ST	Paired associates	− .18	23
44.	Exp. 1-original ST	HE	–	JOL (p)	RL	Paired associates	− .37	40

Table 3 (continued)

Article	Sample	Level	Goal driven	Monitoring (timing)	Effort	Task	*r*	*n*
Undorf and Erdfelder (2013)	Exp. 1-swapped ST	HE	–	JOL (p)	RL	Paired associates	− .37	40
	Exp. 2-original ST	HE	–	JOL (p)	RL	Paired associates	− .31	47
	Exp. 2-swapped ST	HE	–	JOL (p)	RL	Paired associates	− .30	48
	Exp. 3-original ST	HE	–	JOL (p)	RL	Paired associates	− .18	32
	Exp. 3-swapped ST	HE	–	JOL (p)	RL	Paired associates	− .25	33
45. Undorf and Erdfelder (2015)	Exp. 2-high association first (combined)	HE	–	JOL (p)	ST	Paired associates	− .37	20
	Exp. 2-wide range first (combined)	HE	–	JOL (p)	ST	Paired associates	− .42	18
	Study 3	HE	–	JOL (p)	ST	Paired associates	− .36[a]	23
46. Weidemann and Kahana (2016)	One study	HE	–	CR (c)	RL	Paired associates	− .26	171

1–6 = grade 1–6, 7–12 = grade 7–12, *HE* higher education, *CR* confidence rating, *JOL* judgment of learning, *(p)* prospective, *(c)* concurrent, *RL* response latency, *ST* study time, *ME* mental effort. *ME-A* self-agent mental effort rating, *Exp.* experiment, *MC* multiple choice

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

References marked with an asterisk are included in the meta-analysis

*Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. Journal of Experimental Psychology: General, 143(3), 1349–1368. doi:https://doi.org/10.1037/a0035098.

*Ackerman, R., & Koriat, A. (2011). Response latency as a predictor of the accuracy of children's reports. Journal of Experimental Psychology: Applied, 17(4), 406–417. doi:https://doi.org/10.1037/a0025129.

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004.

*Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. Psychonomic Bulletin & Review, 19(6), 1187–1192. doi:https://doi.org/10.3758/s13423-012-0305-z.

*Baars, M., & Wijnia, L. (2018). The relation between task-specific motivational profiles and training of self-regulated learning skills. Learning and Individual Differences, 64, 125–137. doi:https://doi.org/10.1016/j.lindif.2018.05.007.

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

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

*Baars, M., Wijnia, L., & Paas, F. (2017). The association between motivation, affect, and self-regulated learning when solving problems. Frontiers in Psychology, 8:1346. doi:https://doi.org/10.3389/fpsyg.2017.01346.

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

*Ball, B. H., Klein, K. N., & Brewer, G. A. (2014). Processing fluency mediates the influence of perceptual information on monitoring learning of educationally relevant materials. Journal of Experimental Psychology: Applied, 20(4), 336–348. doi:https://doi.org/10.1037/xap0000023.

*Blissett, S., Sibbald, M., Kok, E., & Van Merriënboer, J. (2018). Optimizing self-regulation of performance: is mental effort a cue? Advances in Health Sciences Education, 23, 891–898. doi:https://doi.org/10.1007/s10459-018-9838-x, 5.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: Wiley.

*Burkett, C., & Azevedo, R. (2012). The effect of multimedia discrepancies on metacognitive judgments. Computers in Human Behavior, 28 4, 1276–1285. doi:https://doi.org/10.1016/j.chb.2012.02.011.

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

*Cuevas, H. M., Fiore, S. M., & Oser, R. L. (2002). Scaffolding cognitive and metacognitive processes in low verbal ability learners: Use of diagrams in computer-based training environments. Instructional Science, 30, 433–464. doi:https://doi.org/10.1023/A:1020516301541, 6.

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

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks: Sage Publications.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory* (pp. 23–38). New York: Oxford University Press.

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98. https://doi.org/10.2307/2669529.

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

*Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. Instructional Science, 43, 487–506. doi: https://doi.org/10.1007/s11251-015-9347-8, 4.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186.

*Hoffmann-Biencourt, A., Lockl, K., Schneider, W., Ackerman, R., & Koriat, A. (2010). Self-paced study time as a cue for recall predictions across school age. British Journal of Developmental Psychology, 28, 767–784. doi:https://doi.org/10.1348/026151009X479042, 4.

*Hoogerheide, V., Loyens, S. M. M., & Van Gog. T. (2014). Comparing the effects of worked examples and modeling examples on learning. Computers in Human Behavior, 41, 80–91. doi:https://doi.org/10.1016/j.chb.2014.09.013.

*Jia, X., Li, P., Li, X., Zhang, Y., Cao, W., & Li, W. (2016). The effect of word frequency on judgments of learning: contributions of beliefs and processing fluency. *Frontiers in Psychology, 6,* Article 1995 doi: https://doi.org/10.3389/fpsyg.2015.01995.

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

*Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. Memory & Cognition, 36(2), 416–428. doi:https://doi.org/10.3758/MC.36.2.416.

*Koriat, A. (2018). Agency attributions of mental effort during self-regulated learning. Memory & Cognition, 46, 370–383. doi:https://doi.org/10.3758/s13421-017-0771-7, 3.

*Koriat, A., & Ackerman, R. (2010a). Choice latency as a cue for children's subjective confidence in the correctness of their answers. Developmental Science, 13(3), 441–453. doi:https://doi.org/10.1111/j.1467-7687.2009.00907.x.

*Koriat, A., & Ackerman. R. (2010b). Metacognition and mindreading: judgments of learning for self and other during self-paced study. Consciousness and Cognition, 19, 251–264. doi:https://doi.org/10.1016/j.concog.2009.12.010, 1.

*Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. Journal of Memory and Language, 52(4), 478–492. doi:https://doi.org/10.1016/j.jml.2005.01.001.

Koriat, A., & Nussinson, R. (2009). Attributing study effort to data-driven and goal-driven effects: implications for metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(5), 1338–1343. https://doi.org/10.1037/a0016374.

*Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. Journal of Experimental Psychology: General, 135(1), 36–69. doi:https://doi.org/10.1037/0096-3445.135.1.36.

*Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009a). The memorizing effort heuristic in judgments of learning: a developmental perspective. Journal of Experimental Child Psychology, 102(3), 265–279. doi: https://doi.org/10.1016/j.jecp.2008.10.005.

*Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009b). The easily learned, easily remembered heuristic in children. Cognitive Development, 24(2), 169–182. doi:https://doi.org/10.1016/j.cogdev.2009.01.001.

*Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014a). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: a developmental perspective. Journal of Experimental Psychology: General, 143(1), 386–403. doi:https://doi.org/10.1037/a0031768.

*Koriat, A., Nussinson, R., & Ackerman, R. (2014b). Judgments of learning depend on how learners interpret study effort. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 1624–1637. doi: https://doi.org/10.1037/xlm0000009, 6.

*Kostons, D., & De Koning, B. B. (2017). Does visualization effect monitoring accuracy, restudy choice, and comprehension scores of students in primary education? Contemporary Educational Psychology, 51, 1–10. doi:https://doi.org/10.1016/j.cedpsych.2017.05.001.

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

*Lachner, A., Backfisch, I., Hoogerheide, V., Van Gog, T., & Renkl, A. (2020). Timing matters! Explaining between study phases enhances students' learning. Journal of Educational Psychology, 112(4), 841–853. doi:https://doi.org/10.1037/edu0000396.

*Li, X., Li, W., Li, P., & Jia, A. (2016). The effect of animacy on metamemory. Memory & Cognition, 44, 696–705. doi:https://doi.org/10.3758/s13421-016-0598-7, 5.

Maki, R. H. (1998). Predicting performance on text: delayed versus immediate predictions and tests. *Memory & Cognition, 26*(5), 959–964. https://doi.org/10.3758/BF03201176.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463–477. https://doi.org/10.1016/j.jml.2004.12.001.

*Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? It depends on your beliefs about intelligence. Psychological Science, 22(3), 320–324. doi:https://doi.org/10.1177/0956797610397954.

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

*Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: does it exemplify fluency effects or reflect people's beliefs about memory? Journal of Memory and Language, 70, 1–12. doi:https://doi.org/10.1016/j.jml.2013.09.007.

*Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2016). The effect of identical word pairs on people's metamemory judgments: what are the contributions of processing fluency and beliefs about memory? The Quarterly Journal of Experimental Psychology, 69(4), 781–799. doi:https://doi.org/10.1080/17470218.2015.1058404.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–141). New York: Academic Press.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8.

*Paik, E. S., & Schraw, G. (2013). Learning with animation and illusions of understanding. Journal of Educational Psychology, 105(2), 278–289. doi:https://doi.org/10.1037/a0030281.

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

*Prince, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. Journal of Memory and Language, 94, 177–194. doi:https://doi.org/10.1016/j.jml.2016.12.003.

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction, 51*, 36–46. https://doi.org/10.1016/j.learninstruc.2016.12.002.

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

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

*Roebers, C. M., Mayers, B., Steiner, M., Bayard, N. S., & Van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: a longitudinal study. Developmental Psychology, 55(10), 2077–2089. doi:https://doi.org/10.1037/dev0000776.

*Schnaubert, L., & Bodemer, D. (2017). Prompting and visualising monitoring outcomes: guiding self-regulatory processes with confidence judgments. Learning and Instruction, 49, 251–262. doi:https://doi.org/10.1016/j.learninstruc.2017.03.004.

Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge im children and adolescents. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory* (pp. 491–518). New York: Oxford University Press.

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition in education (pp. 415–428). New York: Routledge.

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

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). New York: Routledge.

*Thompson, V. A., Prowse Turner, J. A., & Pennycook. G. (2011). Intuition, reason, and metacognition. Cognitive Psychology, 63, 107–140. doi:https://doi.org/10.1016/j.cogpsych.2011.06.001, 3.

*Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. Psychonomic Bulletin & Review, 24, 2003–2011. doi:https://doi.org/10.3758/s13423-017-1261-4, 6.

*Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: conclusive evidence for the ease-of-processing hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(5), 1264–1269. doi:https://doi.org/10.1037/a0023719.

*Undorf, M., & Erdfelder, E. (2013). Separation of encoding fluency and item difficulty effects on judgements of learning. The Quarterly Journal of Experimental Psychology, 66(10), 2060–2072. doi:https://doi.org/10.1080/17470218.2013.777751.

*Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: a closer look at the contribution of processing fluency. Memory & Cognition, 43, 647–658. doi:https://doi.org/10.3758/s13421-014-0479-x, 4.

*Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. Royal Society Open Science, 3:150670. doi:https://doi.org/10.1098/rsos.150670, 4.

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

Affiliations

Martine Baars[1] · Lisette Wijnia[2,3] · Anique de Bruin[4] · Fred Paas[1,5]

[1] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2] HZ University of Applied Sciences, Vlissingen, The Netherlands

[3] Erasmus University College, Erasmus University Rotterdam, Rotterdam, The Netherlands

[4] Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands

[5] School of Education/ Early Start, University of Wollongong, Wollongong, Australia