July 2020

A paper in press

Reference:
Scheiter, K., Ackerman, R., & Hoogerheide, V. (in press). Looking at Mental Effort Appraisals through a Metacognitive Lens: Are they Biased? *Educational Psychology Review*. doi:10.1007/s10648-020-09555-9

Looking at mental effort appraisals through a metacognitive lens: Are they biased?

Katharina Scheiter

Leibniz-Institut für Wissensmedien / University of Tübingen

Rakefet Ackerman

Technion—Israel Institute of Technology

Vincent Hoogerheide

Department of Education, Utrecht University

Abstract

A central factor in research guided by the Cognitive Load Theory (CLT) is the mental effort people invest in performing a task. Mental effort is commonly assessed by asking people to report their effort throughout performing learning or problem-solving tasks. Although this measurement is considered reliable and valid in CLT research, metacognitive research provides robust evidence that self-appraisals of performance are often biased. In this review we consider the possibility that mental effort appraisals may also be biased. In particular, we review signs for covariations and mismatches between subjective and objective measures of effort. Our review suggests that subjective and most objective effort measures appear reliable and valid when evaluated in isolation, because they discriminate among tasks of varying complexity. However, not much is known about their mutual correspondence—that is, whether subjective measures covariate with objective measures. Moreover, there is evidence that people utilize heuristic cues when appraising their effort, similar to utilization of heuristic cues underlying metacognitive appraisals of performance. These cues are identified by exposing biases—mismatch in effects of cue variations on appraisals and performance. The review concludes with a research agenda in which we suggest applying the well-established methodologies for studying biases in self-appraisals of performance in metacognitive research to investigating effort appraisals. One promising method could be to determine the covariation of effort appraisals and objective effort measures as an indicator of the resolution of effort appraisals.

Keywords: instructional design; metacognitive monitoring and control; mental effort; self-regulated learning; cognitive load measurement

1. Introduction

While performing daily cognitively demanding tasks, like navigating, designing, decision making, solving problems, and learning new information, people constantly regulate their mental effort, that is, the amount of cognitive resources they allocate to achieve their goal of performing the task. They might invest a lot of effort because the task is highly challenging but carries important implications. Alternatively, they might invest only little effort either because the task appears easy, or because it looks impossible to solve and they give up right away (e.g., when facing a new type of task, a student may announce "we did not learn how to do it" and give up immediately). In all these situations, people must assess their chances of success and ongoing progress, either implicitly or explicitly, and based on this subjective assessment, take regulatory decisions, such as whether to invest additional effort, change a strategy, seek for external help, or just give up on the task.

In the context of educational research, appraisals of effort and task performance are important sources for researchers to understand how people acquire new information and perform on problem-solving tasks. People do so not only by executing cognitive processes but also by reflecting upon and regulating their cognition (Nelson & Narens, 1990). These appraisals can refer to the person performing the task (i.e., self-perceptions), to the task itself, as well as to the ongoing process of learning new information or solving problems (i.e., momentary experiences, see Ackerman, 2019, for a review). Two prominent research areas where student appraisals play a major role are informed by *Cognitive Load Theory* (CLT; Sweller, Ayres, & Kalyuga, 2011; Sweller, van Merriënboer, & Paas, 1998) and *Metacognition* (see Bjork, Dunlosky, & Kornell, 2013; Fiedler, Ackerman, & Scarampi, 2019, for reviews). The latter has been a major theoretical backbone for research on *Self-Regulated Learning* (SRL).

The goal of the present review is to consider whether people's appraisals of the effort that they invest into a task (i.e., their momentary effort experience) are reliable and valid in that they reflect the cognitive resources they allocated[1]. The accuracy of effort appraisals is fundamental from a learning and instruction perspective, because they are pivotal to explaining instructional design effects discussed within the CLT as well as to designing effective interventions and training programs (e.g., Mirza, Agostinho, Tindall-Ford, Paas, & Chandler, 2019; Raaijmakers,

[1] In the remainder of the paper we use the term "mental effort" to refer to the construct as referred to within CLT, whereas we use "effort" when referring to the allocation of cognitive resources more generally speaking.

Baars, Paas, van Merriënboer, & van Gog, 2018; van Gog, Hoogerheide, & van Harsel, 2020). Whereas mental effort appraisals are considered reliable and valid in CLT research, metacognitive research suggests that self-appraisals of performance are often biased. In the present review we adhere to a call by De Bruin and van Merriënboer (2017) and Seufert (2018) to link CLT and metacognition research. In particular, by borrowing well-established methods and concepts for determining the accuracy of subjective appraisals from metacognition research (cf. Ackerman & Thompson, 2017; Bjork et al., 2013, for reviews), we shed light on the question of whether mental effort appraisals may also be biased and suggest methodologies to investigate this question.

2. The role of appraisals within Cognitive Load Theory and Metacognition research

The CLT deals with the cognitive processing induced by the design of educational tasks and how this processing affects students' performance. The manner in which a task is designed can either hinder task performance by overloading the cognitive system with unnecessary cognitive processing (i.e., extraneous cognitive load, Sweller et al., 1998) or foster performance by stimulating cognitive processes relevant for learning (i.e., germane cognitive load, Sweller et al., 1998). In addition, intrinsic cognitive load is caused by features inherent to the task (i.e., task complexity) and the person working on it (i.e., prior knowledge). Thus, cognitive load research focuses on the way task (design) characteristics relative to a person's skills and knowledge level affect cognitive load. That is, cognitive load is the total demands that "performing a particular task imposes on the learner's cognitive system" (Paas, Tuovinen, Tabbers, & van Gerven, 2003, p. 64).[2] As an instructional design theory, the CLT focuses on effects on learning outcomes, while the balance of load types (i.e., extraneous, germane, and intrinsic cognitive load) serves as a mediating variable to explain learning outcomes.

One common way to estimate cognitive load is to ask people to report on the mental effort they invested into task accomplishment. We will elaborate on the relation between cognitive load and mental effort below. A fundamental assumption underlying the use of effort appraisals in CLT research is that they are reliable and valid—reflect the cognitive resources

---

[2] There is an ongoing debate in the CLT research community regarding the types of load and the relation of these load types. Because this discussion is not pertinent to the current paper, we refer to the classical notion of CLT, according to which extraneous, germane, and intrinsic cognitive load are considered to be separable types of cognitive load that when added make up total cognitive load. Readers interested in recently introduced alternatives to this assumption are referred to Sweller, van Merriënboer, and Paas (2019).

allocated to the task—and that effort appraisals are therefore good predictors of learning outcomes.

Metacognitive research[3], on the other hand, emphasizes the fact that task-related cognitive processing and people's judgments of their performance stem from fundamentally different processes. It focuses on how people assess their own chance of being successful throughout performing cognitive tasks (e.g., learning). Nelson and Narens (1990) identified four metacognitive judgments: Ease of Learning judgments take place before learning, reflecting a preliminary impression of task difficulty (or ease); Judgments of Learning during and immediately after learning a particular item (e.g., term, new vocabulary word); Feeling of Knowing is a prediction regarding one's ability to answer a particular knowledge question before searching memory; and Confidence is a retrospective assessment of the chance of a provided answer to be correct. Ackerman and Thompson (2015) paralleled to these four judgment types judgments which take place throughout solving problems: Judgment of Solvability—before solving a problem; Feeling of Rightness—regarding the first solution that comes to mind; intermediate confidence while considering potential solutions; and final confidence—after choosing which solution to provide.

Central to metacognitive research is the massive body of empirical evidence demonstrating biases in people's appraisal of their own chance for success. The source for such biases is well-agreed upon to stem from utilizing heuristic cues as bases for these judgments (cue utilization, Koriat, 1997; see Ackerman, 2019, for a review; e.g., fluency—ease of processing, familiarity, concreteness). These cues are sometimes valid, in that they predict changes in performance (high cue diagnosticity, e.g., Koriat, Ma'ayan, & Nussinson, 2006; Thompson et al., 2013), but also often not tightly associated with performance fluctuations (low cue diagnosticity, e.g., Metcalfe & Finn, 2008; Rabinowitz, Ackerman, Craik, & Hinchley, 1982; Sidi, Shpigelman, Zalmanov, & Ackerman, 2017). Typically, biases in metacognitive judgments are reflected in

[3] We refer in this review to the body of Metacognitive research dealing with monitoring and control of regulatory decisions. It includes Meta-Memory research, dealing with memorizing and knowledge retrieval (e.g., Koriat, 1997), Meta-Comprehension, dealing with reading comprehension tasks (e.g., Thiede, Anderson, & Therriault, 2003), and Meta-Reasoning, dealing with the unique monitoring and control decisions employed when facing reasoning, problem-solving and decision-making tasks (see Ackerman & Thompson, 2015, 2017, for reviews and comparisons among the subdomains). In this review, we refer to metacognitive principles common across these three sub-domains. A large body of metacognitive research focuses on strategies people employ, people's explicit awareness of factors affecting their performance, and for planning their self-regulated learning activities (e.g., Azevedo, Moos, Johnson, & Chauncey, 2010; Winne & Hadwin, 1998). These aspects are out of the focus of the present review.

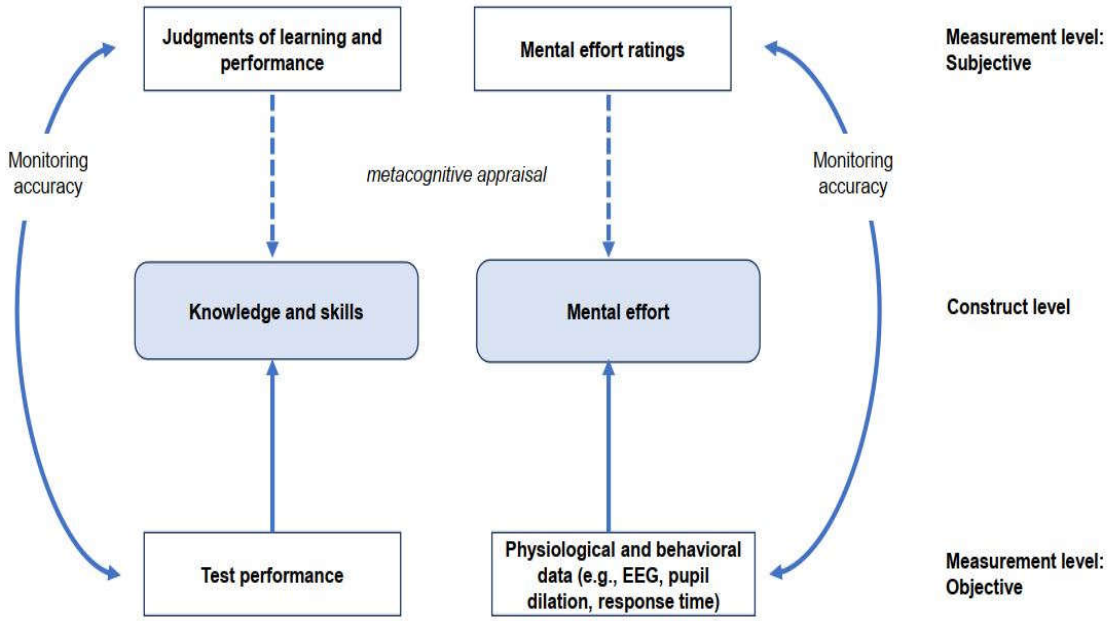
overconfidence, which has been found across samples of children, adolescents, and adults, and in various domains (see Bjork et al., 2013; Destan & Roebers, 2015; Dunning, Heath, & Suls, 2004; Fiedler et al., 2019, for reviews). Moreover, people distinguish only to a moderate extent between their own successes and failures (e.g., Koriat et al., 2006; Thiede et al., 2003). Regarding the timing of metacognitive judgments, it has been shown that the later they are required (i.e., the more experience a person has acquired with the task), the more accurate they are; nevertheless, even retrospective judgments are prone to biases (Leonesio & Nelson, 1990; Siedlecka, Paulewicz, & Wierzchoń, 2016).

Importantly, despite this moderate and sometimes even low appraisal accuracy, metacognitive judgments guide regulatory decisions that people take regarding withholding or providing answers and regarding further investment of effort for restudying, seeking help, or giving up (e.g., Koriat & Ackerman, 2010; Krebs & Roebers, 2010; Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003; Undorf, Livneh, & Ackerman, 2020).

In the present review we propose that effort appraisals are, in fact, a type of metacognitive judgment. Their least common denominator is that they are both subjective as compared to objective measures of effort or performance (cf. Baars, Wijnia, de Bruin, & Paas, 2020). As a working hypothesis we suggest the possibility that also effort appraisals may hence be prone to biases, but nevertheless guide people's actions. This notion has been by-and-large overlooked in the vast research based on CLT, where effort appraisals serve as a tool to explain effects on learning outcomes. We situate our review within the Effort Monitoring and Regulation (EMR) Framework (see De Bruin, Roelle, Baars, & EFG-MRE, 2020). The goal of the framework is to integrate CLT and SRL theory by guiding and stimulating research with respect to three fundamental questions: (1) How do people monitor effort?, (2) How do people regulate effort?, and (3) How do people optimize cognitive load while performing self-regulated learning tasks? The present paper delves into the first question by considering the possibility that effort appraisals used in CLT may not always accord objectively measured effort.

To address this question, we draw an analogy between CLT and metacognitive research, as illustrated in Figure 1. We suggest that both areas of research use subjective measures (appraisals, upper part of the figure) and objective measures (lower part of the figure) to gather insights on constructs related to a person's cognitive state (middle part of the figure). In line with Nelson and Narens (1990), we argue that both judgments of performance as well as of effort are

metacognitive appraisals in that they are based on the way a person thinks about his/her cognition, rather than on the cognition itself. According to metacognitive research, the accuracy of these metacognitive appraisals (i.e., monitoring accuracy) can be determined by examining the relation between subjective and objective measures of the same construct. An important assumption underlying this approach is that objective measures of knowledge and skills, namely test performance, are reliable and valid. Hence, a lack of correspondence between subjective and objective measures (i.e., low monitoring accuracy) can be attributed to biases in metacognitive appraisals.[4] In CLT research, there are only a few studies addressing the reliability and validity of both subjective and objective measures of effort or considering associations among them. The goal of the present review is to discuss whether and how researchers can assess the extent in which judgments of effort are accurate (or biased), by relying on the analogy to metacognitive research.



Note: dashed lines reflect the assumption that the relations between concepts may be subject to metacognitive inference processes; in contrast, solid lines reflect more direct relationships.

Figure 1: Parallels between Metacognitive and Cognitive Load Theory (CLT) concepts

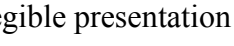
[4] In metacognitive research, an underlying assumption is that test outcomes reflect the knowledge it was designed to assess. In fact, clearly, each test outcome reflects whether an individual expresses in his/her answers the particular knowledge the test examines. For instance, it is possible that a particular test is easy for a person who does not know much beyond the exact knowledge included in the test.

In the remainder of this paper, we first review the literature regarding subjective and objective measures of effort within CLT research. Second, we discuss available research indicating that there might be biases in effort appraisals. Finally, we propose an agenda for future research to investigate the processes that underlie effort appraisals and potentially misguide people's regulatory decisions.

3. Mental effort from a cognitive load perspective

3.1. Definition

Within CLT, the notion of mental effort and cognitive load are closely related. Cognitive load is defined as "a multidimensional construct that represents the load that performing a particular task imposes on the cognitive system of a particular learner" (Paas & van Merriënboer, 1994b, p. 122). It is characterized by two types of factors: those causing cognitive load, namely, task features, individual characteristics, and their interaction; and assessment factors affected by cognitive load (see Paas & van Merriënboer, 1994a,b; Paas, van Merriënboer, & Adam, 1994). With respect to the latter, the CLT literature discusses three types of load indicators, mental load, mental effort, and task performance, which differ in the factors causing them.

First, the term *Mental Load* refers to the demands a task imposes on a person's cognitive resources, most importantly, working memory capacity (i.e., Paas & van Merriënboer, 1994a). The more complex a task, the higher the mental load. This positive relationship between complexity and mental load is seen as independent of a person's characteristics. In the metacognitive literature, this is called data-driven effort (Koriat et al., 2006). Data-driven effort is required by the task in a bottom-up manner, and the person has little role in guiding this effort (assuming sufficient level of motivation for engaging in the task). It means that an easy task (e.g., remembering a short list of words) requires less effort than a more challenging task (e.g., remembering a long list of words). Similarly, a clear presentation of stimuli requires less effort than a less legible presentation (e.g., straight vs. ).

Second, the term *Mental Effort* addresses the human-centered dimension of cognitive load. According to Paas and van Merriënboer (1994a), "mental effort refers to the amount of capacity or resources that is actually allocated to accommodate the task demands" (p. 354). Mental effort is affected by features of the task, individual characteristics, and their interaction. Effort is assumed to result from controlled (as compared to automatic) processing of task-relevant information (cf. Shiffrin & Schneider, 1977). Effort has a motivational aspect to it

(Feldon, Callan, Juth, & Jeong, 2019; Paas, Tuovinen, van Merriënboer, & Darabi, 2005; see also Eitel, Endres, & Renkl, 2020). In the metacognitive literature, this type of effort is called goal-driven effort (Koriat et al., 2006). It is invested by using top-down regulatory processes (e.g., decision to answer "I do not know" or to allocate additional time) guided by the person's self-assessment of performance (Nelson & Narens, 1990; see Ackerman, 2014, for a review). By the metacognitive approach, data-driven effort and goal-driven effort are combined while working on each task item, reflecting the combination between task requirements and people's motivation for success (Koriat et al., 2006). In contrast, CLT is commonly conceptualized as a cognitive theory that does not address student motivation; rather, the theory is assumed to apply to situations where students are highly motivated so that they invest all their available resources into accomplishing a task (Schnotz, Fries, & Horz, 2009). Nevertheless, from a motivational perspective, influences of motivation on effort are likely (e.g., Paas et al., 2005). For instance, a person may decide to allocate fewer resources when task complexity leads to low expectancy of goal achievement or when the task is seen as not sufficiently relevant to warrant high effort investments (i.e., low task value).

According to both theoretical frameworks, *Task Performance* results from the accumulation of all effort types. Accordingly, similar to mental effort, performance is affected by task features, individual characteristics, and their interaction. If a person decides to invest a lot of effort into a complex task, then task performance may be similar to that achieved in a less complex task after investing less effort. Hence, effort moderates the relationship between mental load and task performance.

To conclude, CLT and metacognitive research distinguish between similar dimensions when discussing how cognitive resources are invested into task accomplishment, with mental load corresponding to data-driven effort and mental effort corresponding to goal-driven effort.

3.2. Measuring effort

In line with the multidimensional nature of cognitive load, various measurement methods have been used in CLT research. These methods can be characterized by the *dimension of cognitive load* they aim to measure, whether they are *subjective* or *objective*, and *when* they are measured.

First, measures used under CLT can be more or less explicitly linked to any of the three dimensions discussed above: mental load, mental effort, and task performance. Since the present review is focused on effort appraisals, we do not discuss measures that are unambiguously

related to the two remaining load dimensions, mental load, and task performance. Thus, we do not address dual task performance, which is used for measuring mental load in the main task in the presence (relative to the absence) of a secondary task (e.g., Brünken, Steinbacher, Plass, & Leutner, 2002), or subjective appraisals of task difficulty (Ayres, 2006).

Second, as discussed in detail below, effort measures are typically based on either subjective self-reports or objective behavioral (e.g., response time) and physiological measures (e.g., eye tracking, EEG; see Brünken, Plass, & Leutner, 2003; Sweller et al., 2011, for reviews). The focus of the present review is on the question of whether subjective effort appraisals are biased. To answer this question, it is important to gauge how they relate to objective measurements when performing the same task.

Third, the majority of measures can be administered throughout performing the task (van Gog & Paas, 2008). When effort is assessed during or after performing a task comprised of several items (e.g., exam), it refers to the global knowledge and skills. On the other hand, it can be assessed in proximity to solving each task item, for measuring ongoing fluctuations which differ among items with different characteristics and for examining ongoing changes (e.g., getting experience with the task). In that case, the assessment reflects the ease by which knowledge necessary to solve the task items is retrieved and applied. Knowledge that is easier to access and apply (e.g., related to existing knowledge) has been consolidated better and requires less effort for retrieval and use. Typically, effort measures administered during task performance provide an index of the quality of knowledge acquisition.

3.2.1 Subjective effort measures

The most prominent way for assessing effort is by eliciting subjective appraisals. Using this method relies on the assumption that when asked to introspect on cognitive processes, people can provide a reliable estimate of effort expenditure (Paas, 1992). Accordingly, Paas proposed the effort item, which (with some variability in its formulation and scaling) is used in the majority of CLT research (van Gog & Paas, 2008). It asks people to rate on 9-point Likert scale the mental effort that was invested to accomplish a task (e.g., "*I invested very, very low mental effort … very, very high mental effort"*).

A reanalysis of data reported in Paas (1992) and Paas and van Merriënboer (1994b) is reported in Paas et al. (1994) to deliver insights into the reliability of the mental effort rating. Internal consistency (Cronbach's alpha) of the rating was determined by administering multiple

test problems and asking students to rate their effort after each problem. Cronbach's alpha across these multiple effort ratings was .90 (Paas, 1992) and .84 (Paas & van Merriënboer, 1994b), respectively. However, it has to be kept in mind that internal consistency is actually a reliability measure that can only be meaningfully interpreted when there are multiple items that assess slightly different aspects of the same construct. If items are too similar (or in this case identical), this homogeneity of items yields artificially high reliability scores (cf. reliability-validity dilemma, Lord & Novick, 1968). At the same time, the concept of retest reliability cannot be applied to mental effort, because – in contrast to personality traits – effort is assumed to fluctuate (e.g., because of training effects or fatigue). Accordingly, determining the reliability of the single-item effort appraisal is problematic.

A major benefit of effort self-appraisals is that they are easy to administer, cost-and time-efficient, sensitive to variations in task complexity (see below), and cause only minimal interference with task accomplishment. The self-report question about effort has been used successfully to delineate instructional conditions that yielded corresponding differences in learning outcomes. For instance, students have reported less effort and were more successful when working on completion problems or learning from worked examples compared with solving the same problems without instructional support (Paas, 1992; Van Harsel, Hoogerheide, Verkoeijen, & van Gog, 2019; van Merriënboer, Schuurman, de Croock, & Paas, 2002). Nevertheless, there have also been misalignments where instructional conditions affected effort ratings, which were not accompanied by corresponding changes in task performance though (e.g., van Gog, Paas, & van Merriënboer, 2006).

The exact delivery of the effort appraisal question is hardly discussed in CLT research. One issue to consider is scaling. It is yet an open issue in CLT research whether a person can really discriminate between nine different levels of effort, and from what age on (cf. Sweller, van Merriënboer, & Paas, 2019). Moreover, one may wonder how people interpret the question regarding the amount of effort they invested into the task. Effort can have a negative connotation referring to the strain imposed by a task or it can have a positive connotation referring to the engagement a person shows with the task (Fisher & Oyserman, 2017). These connotations, moreover, differ in the associated perceived agency: the origins of effort as something that is either required by a task or invested based on a voluntary decision of the person. Whereas the original item suggested by Paas (1992) emphasizes the latter, different variants translated into

different languages have been used, which may have different connotations (cf. Sweller et al., 2019). Unfortunately, most studies, including the ones introducing the mental effort item (e.g. Paas, 1992; Paas & van Merriënboer, 1994b), do not report the exact wording of the question about effort. Within CLT research, little is known about how students interpret the effort item and there is no systematic research addressing the impact that differences in wording may have.

Recent metacognitive research suggests that students associate higher effort with study strategies that they perceive as less effective (while in actuality they are not), indicating negative connotations of effort. In particular, in the study by Kirk-Johnson, Galla, and Fraundorf (2019), learners were asked to provide effort appraisals using four items (e.g., 'How tiring was the last exercise'?) and to judge their expected learning outcome after using one of two study strategies (blocked vs. interleaved practice). In addition, learners had to indicate which of the two study strategies they would choose for future learning. Research has shown consistently that interleaved practice is more effective than blocked practice (see Yan, Bjork, & Bjork, 2016, for a review). Although learners gave higher effort appraisals to interleaved practice, as expected, they evaluated interleaved practice as being less effective, and chose it less often as the preferred strategy for future study (see also Carpenter, Endres, & Hui, 2020, regarding the interpretation of higher effort associated with retrieval practice as indicating ineffective learning). It is important to note that at least the question provided as an example in the paper emphasized strenuous task aspects compared with the effort item typically used in CLT research. Nevertheless, this study nicely illustrates that we know very little about how learners interpret effort appraisals.

Koriat, Nussinson, and Ackerman (2014) succeeded in manipulating the way people interpret their effort, to be called for by the task item or stemming from their inner motivation. They achieved this by only changing the phrasing of the question on the appraisal scale to be "1 (The paragraph required little study) to 9 (The paragraph required a great deal of study)" for activating assessment of data-driven effort (or mental load) versus "1 (I chose to invest little study) to 9 (I chose to invest a great deal of study)" for activating assessment of goal-driven effort (or mental effort) (see also Fisher & Oyserman, 2017; Koriat, 2018). This malleability of effort interpretation suggests that people can take the two sources of effort into account and rely on hints in the environment regarding which interpretation to employ when the question is presented.

Another point to consider is that there may be individual differences in how people interpret questions regarding their effort investment. According to Dweck and Leggett (1988), students differ in their implicit theories of intelligence (or mindsets), in that some perceive ability as malleable (growth mindset), whereas others perceive it as fixed. Notably, these differences in mindset have been shown to have an effect, though small, on a variety of motivational and cognitive aspects related to learning and achievement (Peng & Tullis, in press; Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018). In particular, students with a growth mindset see effort as positively related to performance, whereas people with a fixed mindset do not believe that trying harder improves task performance. It is likely that these mindset differences also have an effect on how people assess their effort.

In line with this reasoning, in the context of metacognitive research, people with different mindsets have been shown to differ in how they use encoding fluency (i.e., perceived effort during learning) as a heuristic cue for making judgments of learning. In the study by Miele, Finn, and Molden (2011), students with a fixed mindset were more likely to apply the *easily learned* = *easily remembered* heuristic (Koriat, 2008). Accordingly, they tended to interpret effortful encoding (as evidenced by longer study times for difficult as compared with easy items) as an indication that they had reached the limits of their ability, and, in turn, expected to achieve less successful recall (i.e., low judgments of learning) of items that they found more effortful to learn (see also Kirk-Johnson et al., 2019). This interpretation is in line with data-driven interpretation of effort associated with the bottom-up mental load. Students with a growth mindset, however, interpreted effortful encoding as suggesting greater task engagement. Accordingly, they adapted a top-down, goal-driven effort interpretation and expected better memory (i.e., higher judgments of learning) for items they invested more effort in. Because in this study encoding fluency was a valid cue, mindset also had a systematic effect on the students' calibration. Entity theorists were well calibrated regardless of task difficulty, whereas students with a growth mindset were underconfident for easy items, but overconfident for the most difficult items.

In sum, most CLT research measured effort with the subjective effort item of Paas (1992) or a variant thereof. This effort measure is easy to use and minimally intrusive. Importantly, it is typically sensitive to instructional design variations. However, because it is a single-item appraisal, its reliability is difficult to establish. Moreover, a person's interpretation of effort

assessments may vary depending on the exact formulation and framing of the question as well as on interindividual differences.

3.2.2 Objective effort measures

Effort can also be assessed using objective behavioral and physiological indicators. Time-on-task (i.e., the amount of time a person invests into accomplishing a task) is probably the most intuitive effort measure. Having said so, there are hardly any studies in the context of CLT that have used time-on-task explicitly as a measure of effort. Note that this point demonstrates a stark contrast to metacognitive research, where learning and answering time is discussed extensively as an indicator of effort among adults and children (e.g., Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014; see Baars et al., 2020) and was even manipulated for leading people to believe that the task required more or less effort (Topolinski & Reber, 2010).

Physiological effort measures reflect that when a person modifies his or her engagement with a task, this is typically accompanied by a physiological response. Physiological effort indices that have been discussed in the CLT literature are heart rate variability, skin conductance, brain activity, and pupil dilation (cf. Paas et al., 1994). These measures have their origin in cognitive and neuroscientific research fields, where they serve as indicators for cognitive processing demands. Differently from subjective appraisals, physiological measures hardly interfere with task accomplishment once the apparatus has been prepared for recording and calibrated to the individual. Physiological measures allow for a continuous assessment of effort and are – to varying degrees – highly sensitive to fluctuations of effort (Paas et al., 2003). This fine-tuned sensitivity has been applied in CLT research both for tracking effort fluctuations in shortly-performed task items (e.g., Scharinger, 2018) and as a mean over time when applied for lengthy tasks (e.g., Richter & Scheiter, 2019).

Heart-rate variability reflects changes in the interval between two heart beats and can be measured using either electroencephalogram (ECG) or blood pressure. Notably, there are a number of health-related physical and cognitive conditions that affect interindividual variability and responsiveness of the heart rate and that may occlude intra-individual changes due to task demands (see Forte, Favieri, & Casagrande, 2019).

Electrodermal activity (EDA) or skin conductance measures make use of the fact that the skin becomes a better conductor of electricity with increasing physiological arousal (e.g., Hoogerheide, Renkl, Fiorella, Paas, & van Gog, 2019; Nourbakhsh, Chen, Wang, & Calvo,

2017). EDA consists of phasic EDA, short-term changes (peaks) in response to stimuli, and tonic EDA, longer changes in autonomic arousal (Braithwaite, Watson, Jones, & Rowe, 2013). EDA measurements are affected by room temperature, movement, and bodily actions such as coughing and talking (see Braithwaite et al., 2013). Moreover, about 10% of the population is non-responsive. Nevertheless, research has shown that phasic EDA can be a reliable indicator of cognitive load (e.g., Ruiz, Taib, Shi, Choi, & Chen, 2007).

Brain activity measures can be obtained by using EEG (electroencephalogram) or NIRS (near-infrared spectroscopy). EEG measures immediate electrical activity at various brain locations while a person is performing a task. Changes in theta and alpha waves have been validated as indicators of cognitive processing demands (cf. for reviews Antonenko, Paas, Grabner, & van Gog, 2010; Scharinger, 2018). NIRS is an optical-imaging technology for measuring brain activity used for learning and instruction research (Brucker, Ehlis, Häußinger, Fallgatter, & Gerjets, 2015). Its application is easier than EEG, since it does not require attaching electrodes to a person's head. Rather, it uses infrared light to detect changes in the concentration of oxygenated and de-oxygenated hemoglobin in the blood, which are indicative of cognitive activity. However, whereas EEG can record brain activity in all brain regions, NIRS only works for brain regions located close to the scalp.

Finally, pupil dilation has been used as an objective effort measure. Pupil size can be assessed with most commonly available video-based eye trackers. Starting with the seminal paper by Hess and Polt (1964), it has been shown repeatedly that the pupil dilates with increases in cognitive processing demands (e.g., Richter & Scheiter, 2019; Scharinger, Kammerer, & Gerjets, 2015; Szulewski, Kelton, & Howes, 2017; Zu, Hutson, Loschky, & Rebello, 2019).

To conclude, several objective measures have been used to assess effort during task processing. With technological advances, many technologies have become cheaper, easier to use, provide more information (e.g., multimodal sensors), and allow administration in less controlled environments and with more complex stimuli. The physiological measures diverge in their responsiveness. Whereas changes in heart rate, brain activity measured by EEG, and pupil dilation occur immediately after stimulus onset, brain activity measured by NIRS and tonic EDA reveal changes in effort after a delay. Thus, for the latter measures it can be difficult to synchronize their fluctuations with behaviors of a learner or changes in the stimulus. Furthermore, because there is interindividual variability in physiological responses, all

physiological measures require controlling for an individual's response under baseline conditions. If properly applied, they can also be used in between-subjects designs despite their interindividual variability.

4 Do subjective effort appraisals measure invested effort?

In general, CLT research is based on the assumption that the subjective appraisals of effort are an accurate measure of invested mental effort (Paas, 1992). Before turning to the question of whether this is the case by reviewing the literature, we first take a methodological view on this assumption by discussing how the accuracy of mental effort appraisals could be determined.

4.1 Methodological considerations

From a psychometric perspective, determining whether effort ratings are accurate requires establishing their construct validity, which can be done analytically and/or empirically. From an analytical perspective, a measure is valid if it represents the to-be-measured construct in a comprehensive (content validity) and plausible (face validity) way. From an empirical perspective, the measure has criterion validity if it correlates highly with another measure of the same construct for which validity has already been established (concurrent validity) and/or if it predicts an external criterion based on theory or previous research (predictive validity).

As done with metacognitive assessments of success, concurrent validity of subjective effort appraisals should be established by relating them to objective measures of effort. Research on metacognition typically examines concurrent validity of metacognitive assessment of success by investigating to which degree self-assessment of performance is related to actual task outcome. High correspondence suggests that people are accurate in monitoring their knowledge and performance, whereas low correspondence suggests low monitoring accuracy (see Figure 1). There are two prominent measures of monitoring accuracy, *calibration* and *resolution* (see Ackerman, Parush, Nassar, & Shtub, 2016; De Bruin & van Merriënboer, 2017, for reviews).

An examination of *calibration* distinguishes between a well-calibrated assessment, a positive bias (overconfidence), and a negative bias (underconfidence) (e.g., Finn & Metcalfe, 2014; Sidi et al., 2017). Calculating calibration regarding assessments of success and task outcomes is relatively straightforward, because both can use the same metric (e.g., % correct). That is, let us consider a case in which Jim's mean confidence judgment across answers reflects being 80% sure about his success. If indeed 80% of his answers are correct, Jim is well-calibrated. If in fact only 65% of his answers are correct, he is overconfident.

Applying calibration as an accuracy measure of effort appraisals means that a person is biased if s/he reports either higher or lower effort than was actually invested as indicated by an objective measure. There are at least two challenges here. First, calculating calibration requires an objective standard to which the subjective appraisal is compared to. As mentioned earlier, for metacognitive research, performance is what people are asked to assess and thus it is reliable by definition. On the other hand, the validity of objective measures of effort should be considered in depth, as there is no concrete and well-agreed upon objective measure which people are requested to adjust their appraisals to. In particular, objective effort measures only indicate that there is a change in physiological arousal (alertness). Moreover, physiological arousal is not specific to cognitive processes, but is also associated with emotional responses, for instance. Accordingly, the interpretation of physiological effort measures is affected by what has been called the reverse inference problem in neuroscience (Poldrack, 2001). This neuroscience problem arises when trying to identify which brain activation measure can be considered an indicator of a certain mental process. Such identification can be done only if activation of this brain area is linked to this specific process and does not occur for other processes as well. Following the same inference logic, the interpretation of physiological arousal as effort requires a sound understanding of the task and its requirements as well as triangulating it with other indicators. From this perspective, the validity of objective effort measures can be determined only by accumulating evidence that the measures fluctuate as one would expect if they indeed reflected effort. For instance, effort measures should react to manipulations of task complexity, as discussed in Section 4.3 below. Importantly, this challenge is true for the validation of any psychological construct. A second challenge with applying the calibration logic to effort appraisals is that objective measures do not have an absolute minimum and maximum score and clear meaning to any terminology people tend to use in daily language. This limitation precludes using calibration for measuring the accuracy of effort appraisals.

Resolution is the second monitoring accuracy measure commonly used in metacognitive research. Resolution is calculated as a within-participant correlation between judgment (e.g., confidence) and success (e.g., answer correctness, yes/no) across items (e.g., a question) within a task (e.g., an exam). It is reported as a numerical value on the continuum between -1 and 1 and statistically compared to zero for assessing its strength. This correlation reflects the extent of correspondence between judgment and success in each item—that is, whether higher judgments

(e.g., confidence ratings) are attached to correct responses rather than to wrong ones. Good resolution allows the person to make effective choices by distinguishing among those items where no further effort should be invested and those that should be reconsidered. For instance, when learning towards an exam, the student must choose which items to restudy on the day before the exam in order to make the best out of the remaining time. Metacognitive studies have shown that conditions that support better resolution yield better restudy choices than conditions with weaker resolution (Thiede et al., 2003).

Resolution is highly relevant for effort measures, as it is relative in nature, and considers whether fluctuations in one measure are reflected in the other measure and vice versa. Applying resolution to effort ratings requires administering subjective and objective effort measures simultaneously for several items, allowing for statistically robust enough within-participant correlations. In most metacognitive studies resolution is applied with at least ten items per participant (e.g., Ackerman, 2014; Koriat et al., 2006; Metcalfe & Finn, 2008). However, there are studies with lengthy reading comprehension tasks which applied it with only six items that varied in difficulty (e.g., Thiede et al., 2003).

Moreover, under the assumption that the motivation to solve a series of tasks is kept constant, a person should invest more effort into the task items that are more complex compared with easier ones. This is at least true for tasks that do not become so hard that people give up on them. In the psychometric literature a measures' ability to discriminate among situations is called sensitivity. We will refer to sensitivity as a proxy for resolution as determined in metacognitive research. Why a proxy? As mentioned above, in metacognitive research, the subjective appraisal associated with each item and the objective success measure refer to the same well-agreed upon construct (cf. Figure 1).

In Section 4.2 below, we review CLT studies aimed at determining whether effort measures can discriminate between items with different characteristics, namely, their complexity, as a proxy for resolution. In Section 4.3, we discuss the few CLT studies that have applied both types of effort measures within one study and that hence in principle would allow to determine resolution the way it is done in metacognition research, namely as the degree of co-variation between subjective and objective effort measures.

Finally, predictive validity of subjective effort appraisals can be determined by relating them to other constructs that are assumed to be strongly related to effort. Most importantly, in

CLT research, effort is assumed to be causally related and hence to predict performance. Thus, in experimental studies, predictive validity may be accounted as biased when effort appraisals and performance do not align, that is, if one measure shows variations while the other does not. Moreover, such a finding supports dissociation between subjective and objective measures of effort and indicates that there is room to understand what aspects of the objective measure are ignored and what heuristic cues within the situation are taken into account when assessing effort, but in fact do not affect success in the task.

4.2 Sensitivity to variations in task complexity—a proxy for resolution

In this section we review findings that suggest that effort measures are sensitive to manipulations of task complexity for the following purposes: First, if subjective effort appraisals allowed to discriminate between items of varying task complexity, this would constitute a proxy for resolution in metacognition research. Second, if objective measures were sensitive to task complexity variations, this would point towards their validity as effort measures. Thus, triangulating object and subjective measures with complexity may be useful as a proxy for resolution in metacognitive research. We are not aware of any study that considered full triangulation of these measures. However, a few studies associated either subjective appraisals or objective indications of effort with task complexity.

When considering subjective effort appraisals, in several studies students were asked to work on multiple task items for which task complexity varied within participants, which accords with the requirements for calculating resolution. These studies provide empirical evidence that variations in task complexity are reflected in the task-related subjective appraisals of effort. Several studies have found effort ratings to be higher for complex tasks than for simpler ones (e.g., Paas et al., 1994; Schmeck, Opfermann, van Gog, Paas, & Leutner, 2015; van Gog, Kirschner, Kester, & Paas, 2012).

Complementing these studies, several studies employing physiological measures have shown objective indicators of effort to be sensitive to variations in task complexity. In particular, EEG studies have revealed decreases in EEG alpha frequency band power at parietal electrodes and increases in frontal-central EEG theta frequency band power with increasing task complexity (see Scharinger, 2018, for a review). Moreover, Aghajani, Garbey, and Omurtag (2017) found evidence that combining EEG parameters with markers simultaneously obtained with NIRS significantly improved discrimination between tasks complexity levels compared with using

EEG alone. Using eye tracking, Scharinger et al. (2015) demonstrated increases in pupil dilation as a function of increasing task complexity. Finally, Nourbakhsh et al. (2017) used learners' skin response (EDA) to distinguish between arithmetic learning tasks of different complexity. Based on a number of EDA features they could classify four levels of cognitive load with up to 80% accuracy. In contrast, Haapalainen, Kim, Forlizzi, and Dey (2010) as well as Larmuseau, Vanneste, Cornelis, Desmet, and Depaepe (2019) found no differences in EDA as a function of task complexity. Similarly, Paas et al. (1994) were unable to find evidence that heart rate variability could detect variations in task difficulty.

To conclude, the few studies that (indirectly) investigated resolution of effort measures with several tasks within participants reveal that both subjective and objective measures are quite sensitive to varying complexity. The only exception constitutes skin conductance and heart rate variability measures for which conflicting results were obtained. In contrast to metacognitive research, the degree of resolution is not quantified in terms of within-participant correlation. That is, while in metacognitive research the extent of within-participant correlation between metacognitive judgments and performance is determined and statistically examined, in CLT research researchers have looked at the general alignment by determining whether more complex tasks also yield higher effort appraisals. This approach leaves open how well effort appraisals are aligned with variations in effort as indicated by objective measures. It also leaves open questions about potential biasing factors which generate more fine-tuned variations for identifying conditions, individual characteristics, global task and situation characteristics, and item characteristics that support better sensitivity to effort variations than others.

4.3 Direct comparisons between subjective and objective effort measures

In the previous section we focused on studies analyzing either subjective or objective effort measures as a function of task complexity. This association may be interpreted as a proxy for the resolution measure in metacognitive research. In this section, we review studies that compared results obtained from *simultaneously* applying subjective appraisals and objective effort measures within one study. Notably, none of the studies attempted to analyze the correspondence between the measures by calculating resolution on a continuum. Moreover, some of the studies used mental load ratings instead of, or in addition to, effort appraisals.

Paas, van Merriënboer, and Adam (1994) compared fluctuations in the subjective effort ratings, as suggested by Paas (1992), and heart rate variability as a function of cognitive

demands elicited by several task designs. While the subjective effort ratings reflected differences in task design, the analysis of heart-rate variability revealed only differences between active and resting states, but reflected no more fine-grained sensitivity (see also Nickel & Nachreiner, 2003).

Antonenko and Niederhauser (2010) used EEG to investigate cognitive demands while learners studied two versions of the same hypertext article. They found differences in alpha and theta-activity measures: A hypertext whose links were augmented with leads (i.e., previews of the target page) yielded better learning outcomes and less effort investment as indicated by the EEG measures than a version without leads. Notably, though, subjective effort appraisals did not reveal a difference between the hypertext versions, suggesting that the subjective measure was less sensitive to the processing differences induced by the hypertext design.

Korbach, Brünken, and Park (2017) applied subjective and objective effort measures in the presence of seductive details. The addition of interesting but irrelevant details to the multimedia instruction decreased learning outcomes as expected. Eye-tracking analyses showed that materials were processed differently when seductive details were present. High task difficulty appraisals (i.e., ratings of mental load) were linked to lower performance. In contrast, effort appraisals and pupil dilation were unrelated to learning outcome. This finding highlights potential differences between subjective appraisals of difficulty and of effort. It suggests that there is better correspondence between subjective and objective measures of effort than between difficulty appraisals and objective measures of effort, at least with regard to pupil dilation. Future research is called to examine this difference further.

Richter and Scheiter (2019) documented signs for a dissociation between subjective appraisals and objective effort measures. They studied the effects of signals to highlight text-picture correspondences (e.g., color coding) for students with either low or high prior knowledge. Consistent with previous research (Richter, Scheiter, & Eitel, 2016, 2018), students with low prior knowledge, but not those with high prior knowledge, showed better learning outcomes when text-picture signals had been included rather than excluded in the multimedia materials. A corresponding, albeit weak interaction was observed for pupil size suggesting that low-prior knowledge students had larger pupil sizes in the condition without signals (hence investing more effort) than in the condition with signals. No differences were observed for

subjective appraisals of task difficulty. Unfortunately, subjective effort measures were not used in this study.

To conclude, the few studies that collected both subjective appraisals and objective effort measures suggest that there are dissociations among them. Interestingly, the authors of the aforementioned studies interpret it either as an issue concerning the validity of the subjective appraisals (e.g., Antonenko & Niederhauser, 2010) or of the objective measures (Paas et al., 1994). At this point, it seems that measurement issues regarding both indices contribute to the different effects on the various indices. Both the subjective and objective measures do not show the pattern of results that one would expect based on performance, thereby suggesting on limited predictive validity for both, and also on the merit of studying the underlying mechanism for these complex associations among all sorts of measures. Notably, in contrast to metacognitive research, statistical measures quantifying their degree of association are not provided. Rather, interpretations are made based only on the presence/absence of differences between experimental conditions in the effort measures and their correspondence. Hence, no conclusions can be drawn as to the degree to which the measures do (not) align beyond the mismatch of effects across measures.

4.4 Heuristic cues underlying effort appraisals

As mentioned briefly above, in metacognitive research it is well-established that people do not have access to their actual knowledge. They assess their chance of success by utilizing heuristic cues which are both experience-based (gut feeling, mainly based on fluency) and theory-based (based on beliefs and lay theories; Koriat, 1997). Referring to effort appraisals as metacognitive judgments, brings to the fore the heuristic cues people may utilize to infer their effort assuming that they cannot sense reliably their effort. Heuristic cues are valid as long as they have high cue diagnosticity (Koriat, 1997). Appraisal biases occur when people make use cues with low diagnosticity. As a result, the methodology for exposing utilization of heuristic cues is based on identifying factors that affect metacognitive judgments while having no effect (or even an opposite effect) on objective measures (see Ackerman, 2019, for a review).

Within CLT, an interesting question related to the accuracy of effort appraisals refers to the anchors people use when judging effort to be high or low. Xie and Salvendy (2000) discussed how cognitive load fluctuates during task performance, while we in the following apply their reasoning to mental effort (see also Paas et al., 2003). Following Xie and Salvendy (2000), there

are several ways of conceptualizing effort variations during performing the task. There is the effort that fluctuates during performing a task on a moment-to-moment basis (instantaneous load), accumulated load is the total amount of effort for one task, average effort is the effort experienced during one task on average, and peak effort is the maximum effort experienced during performing the task. When asking people to provide a effort appraisal after working on a task, it is unclear what type of anchor is used as a reference point when making the appraisal. This unique aspect calls attention to the timing and frequency of effort appraisals.

Van Gog et al. (2012) investigated how immediate effort ratings provided after each task item relate to a single rating provided at the end of the task sequence (i.e., a delayed rating). They found that single delayed estimates of effort were higher than the average of all the immediate ratings. Nevertheless, the delayed rating was not as high as the effort students reported to have invested in the most complex task. When the complexity of the items within a task sequence was kept homogenous, a delayed rating was higher than average ratings for task sequences containing only complex problems, but not for those containing only simple problems. Another experiment showed that informing students beforehand that they would have to provide effort ratings resulted in lower delayed ratings. These findings suggest that itemized appraisals are more sensitive to effort fluctuations. This is also in line with metacognitive research in which itemized judgments are used for examining judgment reliability and heuristic cues that underlie them. Summative (or aggregated) judgments are accounted as a different kind of judgment. Such summative judgments can be used for examining individual differences in correlation with other global measures, as done regarding psychometric tests (e.g. Dentakos, Saoud, Ackerman, & Toplak, 2019). Comparisons between itemized and summative metacognitive judgments of success are scarce (e.g., Dentakos et al., 2019; Rivers, Dunlosky, & Joynes, 2019).

Notably, van Gog et al. (2012) interpreted the results of their experiments as suggesting that monitoring effort requires cognitive resources and is not done spontaneously, which is problematic especially in the case of complex problems. This conclusion means that future empirical studies focused on effort appraisals should have a control group which does not rate their effort, for exposing potential reactivity of the appraisals on other measures. In metacognition, the issue of judgment reactivity is under a lively debate in the recent years because of mixed results, with no satisfactory theoretical understanding regarding the conditions that make judgments reactive (see Double, Birney, & Walker, 2018, for a meta-analysis

regarding judgments of learning; see Double & Birney, 2019, for a review of effects on problem solving).

Schmeck and colleagues (2015) followed up on van Gog et al. (2012) by looking more closely into what predicts delayed effort ratings. They conducted two studies in which they investigated the role of task complexity on effort, task difficulty, interest, and motivation ratings. Students were asked to solve six reasoning problems. The results revealed that irrespective of how tasks were sequenced, delayed effort and difficulty ratings were higher than the average ratings, thereby replicating findings from van Gog et al. (2012). No corresponding differences between average and delayed ratings were found for either interest or motivation. Immediate ratings of more complex problems were best suited to predict delayed ratings, at least when presented in a simple-to-complex sequence, whereas the results for a complex-to-simple sequence were less clear. This again suggests that there may be memory biases involved in delayed subjective ratings (cf. Tsang & Wilson, 1997). That is, experienced high peaks in effort and task difficulty as well as serial position both seem to contribute to later appraisals given after the end of the experience (cf. peak-end effect, Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). Research has indeed shown that sequencing learning experiences to end with the most pleasant (i.e., least effortful) part improved the valence of children's memories of learning vocabulary or receiving peer feedback (e.g., Hoogerheide & Paas, 2012; Hoogerheide, Vink, Finn, Raes, & Paas, 2018). Ending on a pleasurable note also influenced children's behavioral preferences (e.g., which task to repeat), thereby affecting their regulation of future study behavior.

Another way of assessing the accuracy of subjective effort appraisals is to investigate whether they are affected by experimental manipulations that are unrelated to the effort invested. Raaijmakers, Baars, Schaap, Paas, and van Gog (2017) had students work on problem-solving tasks. Before they rated their effort at the end of the task sequence, in some conditions they were given (fake) feedback regarding their task performance, which was either positive or negative. In three experiments with different tasks and study populations, the authors showed that students rated their effort higher after negative feedback than after positive feedback. Effort ratings in the no-feedback conditions fell in between. These findings show that feedback affects perceptions of effort investment, thereby serving as an invalid cue for estimating one's effort. Moreover, they

suggest that people tend to associate high effort with ineffective learning (cf. Kirk-Johnson et al., 2019).

Taken together, there are a few studies that identified heuristic cues that people use to base their effort appraisals on. The existing studies suggest that subjective effort appraisals reflect more than just the effort invested in a task, as the effort measurement is influenced by factors that do not have anything to do with the tasks themselves, such as provision of feedback or task position. This suggests that in their effort appraisals learners also make use of cues that are invalid in terms of effort expenditure.

5 Directions for future research

In this review we raise the possibility that effort appraisals do not necessarily reliably reflect the invested effort. Following metacognitive research, we assume that effort appraisals are nevertheless used as bases for people's regulatory decisions and are thus important for understanding learning. Clearly, questioning the accuracy of effort appraisals is a trigger for numerous future research directions. Inspirations for such examination can be taken from seminal studies that provide evidence for such regulatory role in metacognitive research (e.g., Metcalfe & Finn, 2008; Thiede et al., 2003).

Our approach to this possibility was to review relevant empirical findings obtained within the research tradition of CLT and enrich this review with insights obtained from metacognitive research, where subjective judgments of one's own performance have been shown to be informed by various heuristic cues and thus potentially biased. Relating these two research fields allowed us to identify gaps in the literature. In the following paragraphs, we suggest future research directions with regard to three aspects: (a) characteristics of effort appraisals within CLT and their interpretation, (b) measuring resolution of effort measures, and (c) identifying heuristic cues that underlie and might bias effort appraisals.

5.1 Characteristics of effort appraisals within CLT and their interpretation

CLT research has deployed a variety of appraisals that differ in focus (task difficulty vs. effort), wording (effort as engagement vs. strain), scaling (5-, 7-, or 9-point Likert scale), timing of administration, and target detail level (itemized vs. summative). We list here potential research directions that will contribute to understanding and improving elicitation procedures of effort appraisals.

First, there is the issue of scaling discussed above, questioning whether people are sensitive enough to their effort for using Likert scales to rate it. Raising this point calls for a systematic investigation of the instruments and their administration. Such research should involve both quantitative analyses, as discussed above, and qualitative analyses (e.g., think aloud studies) for addressing how people reason and which cues they use when making effort appraisals. Respective quantitative and qualitative studies should also take into account individual characteristics, such as a person's implicit theories of ability (e.g., intelligence) and performance, prior knowledge, and age, which are likely to influence how a person interprets effort and discriminates between different levels of effort.

Second, our review suggests that administration aspects may affect effort appraisals in various ways. For instance, as reviewed above, priming of mindset (Fisher & Oyserman, 2017) and the wording of the question by which effort appraisals are elicited (Koriat, Nussinson, et al., 2014) were found to affect people's perceptions of effort as being either data-driven or goal-driven. These findings demonstrate the malleability of effort appraisals and call researchers to pay attention to these seemingly negligible details which in fact may cause predictable biases.

Third, as mentioned above, calibration, in terms of under- and overconfidence, cannot be calculated without common units agreed between the participants and the researchers for both appraisals and objective measures. As a bypass, in an attempt to allow researchers to use methodologies they typically use for their CLT-inspired studies, we suggest considering standardizing both subjective and objective measures into z-scores. However, the use of z-scores should be done with awareness that the participants cannot have this relativity in mind. While relative, rather than an absolute measure as calibration should be, comparing relative subjective and objective measures across conditions might be useful for identifying conditions in which people tend to be biased upwards or downwards. This methodology is expected to be particularly useful when the factors have a differential effect on the two types of measures (one goes up while the other goes down or unaffected). Notably, in order to calculate z-score and/or resolution, it is recommended to use effort appraisals scales with at least 9 levels, with more being better. This seems to contradict the issue of scaling and sensitivity of people to fine-tuned differences. However, in metacognitive research researchers typically use 0-100% scales. Some people do anchor to a particular level and move around it, but still research shows that most people do use the entire scale and their judgments reflect calibration and resolution variations

across conditions and populations (e.g., Dentakos et al., 2019; Koriat, Ackerman, et al., 2014; Metcalfe & Finn, 2008).

5.2 Resolution of effort measures

This review suggests that subjective and objective indicators of effort are sensitive to variations in task complexity, which can be interpreted as a proxy for good resolution of these measures. However, there are a number of caveats to this interpretation. First, studies investigating this issue are scarce and have also produced conflicting results as in the case of skin conductance and heart rate variability measures. Second, these few studies did not quantify the degree of resolution or investigate whether measures differ in their sensitivity to subtle changes in complexity. Here we call for future studies to control for potential confounding factors, such as individual or contextual factors, and to use sufficiently large item collections and sample sizes. These studies should aim to deploy a variety of task complexity levels for allowing to investigate sensitivity. Finally, more than just one effort measure should be deployed to also allow for judging the relative sensitivity of these measures. Moreover, using several measures allows investigating the co-variation of multiple effort measures across a series of multiple tasks, which could potentially provide a reliable measure of appraisals' accuracy (see below).

A central challenge concerns measuring effort objectively. The physiological effort measures reviewed above are all known to be prone to being affected by individual and contextual characteristics. Thus, they need to be carefully administered to reduce potential artefacts. In particular, there is a need to assess baseline measures for all of them so that effects on the effort indicator can be determined as a change relative to the baseline. Moreover, it is important to keep in mind that objective measures are just an approximation regarding the true value of mental effort invested into a task. The only way to validate these measures is to accumulate evidence that they behave in line with what would be expected from a measure assessing effort, since there is no other way to generate ground truth for any psychological construct. However, we believe that it is important to keep in mind that the same critique also applies to metacognitive research. Performance observed in a learning outcome test is, if the test is well designed, a highly valid instrument for assessing a learner's mental representation, but performance in this test is not identical to the learner's knowledge and skills. Thus, in both research fields we can only hope for approximating true values as a basis for determining the accuracy of subjective appraisals. The difference between the domains lies in using terminology

participants can understand. In metacognitive research, we can ask participants to assess concretely how many questions they answered correctly, or what is the chance of each of their answers to be correct and sum up these appraisals across the exam. When asking participants to assess effort, the terminology is abstract. This is a fundamental difference that no methodology can overcome. Thus, CLT research plans should be even more careful in interpreting the results and attempt to triangulate more measures to increase researchers' confidence in the measures' validity.

Importantly, research designs aiming to look into the correspondence between subjective and objective effort measures should be planned rigorously, while employing critical thinking in light of considerations we brought in this paper and probably others. For instance, subjective appraisals and physiological measures should be collected at the same time points during task performance. This practice should reduce differences in the anchors that people use to make various appraisals (Tsang & Wilson, 1997; Xie & Salvendy, 2000). In particular, measuring subjective and objective effort multiple times during performing a lengthy or complex task will allow investigating the co-variation of the indicators across time, as done in metacognitive research when using resolution for assessing judgment accuracy.

5.3 Identifying heuristic cues for self-assessment of effort

Metacognitive research has provided a plethora of insights regarding the heuristic cues that people utilize when evaluating their own cognition. As mentioned above, the central method for identifying heuristic cues people take (or fail to take) into account in judgment of their performance is to generate dissociation between judgment and actual success in the task in their response to variations in a potentially misleading factor. Our review pointed to initial indications for such dissociations within the CLT literature. This is clearly only a hint that should call attention of researchers to a potential "gold mine" of research topics. Clearly, this research agenda depends heavily on having acceptable indications for effort. However, even using the current methods for pointing to factors that affect differently effort appraisals and objective indications of effort is highly valuable for pushing our theoretical understanding forward. In particular, Ackerman (2019) presented a taxonomy for considering heuristic cues and explained methodologies for doing so in Meta-Reasoning research—metacognitive processes associated with reasoning, problem solving, and decision making. She pointed out cues at three levels, self-perception of the individual, task-related factors, and ongoing experiences people have while

performing each task item. Overall, people tend to underestimate in their metacognitive judgments effects of individual differences and task-related characteristics on performance, and reflect more upon perceived difficulty variations across task items (e.g., Rabinowitz et al., 1982). This analysis highlights that within CLT studies we should distinguish effects on effort in cases of between-participant variations (e.g., Antonenko & Niederhauser, 2010) and within-participant variations (e.g., Paas et al., 1994). Future studies are called for to consider which level of cues are taken into account more strongly in effort appraisals and what can we learn from it regarding conditions that support valid effort appraisal. Confirmatory evidence regarding biases in effort appraisals would come from experiments that reveal effects of heuristic cues on subjective appraisals even though they should not impact effort, whereas objective measures remain unaffected, and vice versa. Such double-dissociation research does not exist so far.

Moreover, subjective sense of effort is a key item-level experience-based heuristic cue in metacognitive research. Typically, this cue is operationalized by response time, as reviewed above and by Baars et al. (2020). Following this review, CLT researchers may find it useful to add response time to their toolbox of objective measures of effort (see Figure 1). Similarly, metacognitive researchers should become aware that response time is clearly not the sole basis for subjective effort experiences, as it seems plausible that people have varying levels of perceived effort in the same invested time and vice versa. Thus, understanding better the mechanisms underlying subjective effort appraisals will also inform back metacognitive theorizing and provide it with better research tools for examining judgments of performance and the following regulatory decisions. Interestingly, associations between subjective effort ratings and metacognitive judgments have shown to be stronger than those between objective effort measures and metacognitive judgments (Baars et al., 2020), suggesting that the prior may share a common denominator, namely, the subjective appraisal process, that eventually yields higher correspondence between them. Accordingly, there may be common mechanisms underlying both types of appraisals that are interesting for coordinated empirical investigations and theory building alike.

6. Conclusions

The effort people invest in performing cognitive tasks is a driving factor of human's achievements. The possibility that subjective appraisals of effort might be based on heuristic

cues and thus might not be tightly associated with the cognitive resources allocated to accomplish the task demands may be astonishing for some researchers and absolutely not surprising for others.

Importantly, we do not claim in the present review that there already is a definite answer to the question of whether effort ratings are biased. Rather, we highlighted in our review that there are initial indications in this direction, although clearly not enough to support answering the question. Moreover, addressing the question of whether effort appraisals are biased is not a trivial task. The methodological and conceptual toolbox borrowed from metacognition research can be useful to refine research on mental effort appraisals as demonstrated above; however, like all analogies, there are limitations to its applicability. A major contribution of this review is in delineating a research agenda that relies on linking CLT research with metacognitive research. With this research agenda, we would like to take the opportunity to raise the flag and call researchers from both research domains to address the challenges highlighted in this review either separately, or, even better, collaboratively for maximum benefit to learning sciences. Importantly, we do not expect an answer to the question of whether effort appraisals are biased any time soon. Thus, we fully agree with De Bruin and van Merriënboer (2017, p. 2): “It has cost 25 years to bring CLT and SRL research to where they are now. Development of a combined theoretical approach and research paradigm and generation of robust insights will probably take at least a similar amount of time.” Thus, we, as a community of researchers, better pave our ways within this forest sooner than later.

References

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

Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psychological Topics, 28*, 1-20. https://psycnet.apa.org/doi/10.31820/pt.28.1.1

Ackerman, R., Parush, A., Nassar, F., & Shtub, A. (2016). Metacognition and system usability: Incorporating metacognitive research paradigm into usability testing. *Computers in Human Behavior*, *54*, 101-113. https://doi.org/10.1016/j.chb.2015.07.041

Ackerman, R., & Thompson, V. (2015). Meta-Reasoning: What can we learn from meta-memory. In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory* (pp. 164-182). Hove, UK: Psychology Press.

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

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring mental workload with EEG+ fNIRS. *Frontiers in Human Neuroscience*, *11*, 359. https://doi.org/10.3389/fnhum.2017.00359

Antonenko, P. D., & Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior*, *26*, 140-150. https://doi.org/10.1016/j.chb.2009.10.014

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*, 425-438. https://doi.org/10.1007/s10648-010-9130-y

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*, 389-400. https://doi.org/10.1016/j.learninstruc.2006.09.001

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*, 210–223. https://doi.org/10.1080/00461520.2010.515934

Baars, M., Wijnia, L., de Bruin, A., Paas, F. (2020). The relation between students' effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review.*

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444. https://doi.org/10.1146/annurev-psych-113011-143823

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analyzing electrodermal activity (EDA) and skin conductance responses (SCRs) for psychological experiments. *Psychophysiology, 49*, 1017–1034.

Brucker, B., Ehlis, A. C., Häußinger, F. B., Fallgatter, A. J., & Gerjets, P. (2015). Watching corresponding gestures facilitates learning with animations by activating human mirror-neurons: An fNIRS study. *Learning and Instruction*, *36*, 27-37. https://doi.org/10.1016/j.learninstruc.2014.11.003

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*, 53-61. https://doi.org/10.1207/S15326985EP3801_7

Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49*, 109-119. https://doi.org/10.1027//1618-3169.49.2.109

De Bruin, A. B. H., Roelle, J., Baars, M., & EFG-MRE (2020). Synthesizing Cognitive Load and Self-Regulation Theory: A theoretical framework and research agenda. *Educational Psychology Review.*

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

Carpenter, S., Endres, T., & Hui, L. (2020). Students' use of retrieval in self-regulated learning: Implications for monitoring and regulating effortful learning experiences. *Educational Psychology Review.*

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

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*, 347-374. https://doi.org/10.1007/s11409-014-9133-z

Double, K. S., & Birney, D. P. (2019). Do confidence ratings prime confidence? *Psychonomic Bulletin & Review*, *26*, 1035-1042. https://doi.org/10.3758/s13423-018-1553-3

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*, 741-750. https://doi.org/10.1080/09658211.2017.1404111

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273. https://doi.org/10.1037/0033-295x.95.2.256

Eitel, A., Endres, T., & Renkl, A. (2020). Self-Management as a bridge between cognitive load and self-regulated learning: The illustrative case of seductive details. *Educational Psychology Review*.

Feldon, D. F., Callan, G., Juth, S., & Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review*, 319–337. https://doi.org/10.1007/s10648-019-09464-6

Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. In R. J. Sternberg & J. Funke (Eds.). *Introduction to the Psychology of Human Thought* (pp. 89-111). Heidelberg: Heidelberg University Publishing.

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*, 1-9. https://doi.org/10.1016/j.learninstruc.2014.01.001

Fisher, O., & Oyserman, D. (2017). Assessing interpretations of experienced ease and difficulty as motivational constructs. *Motivation Science*, *3*, 133-163. https://doi.org/10.1037/mot0000055

Forte, G., Favieri, F., & Casagrande, M. (2019). Heart rate variability and cognitive function: A systematic review. *Frontiers in Neuroscience*, *13*:710. https://doi.org/10.3389/fnins.2019.00710

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 301-310). ACM. https://doi.org/10.1145/1864349.1864395

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143(3611)*, 1190-1192. https://doi.org/ 10.1126/science.143.3611.1190

Hoogerheide, V., & Paas, F. (2012). Remembered utility of unpleasant and pleasant learning experiences: Is all well that ends well? *Applied Cognitive Psychology*, *26*, 887-894. https://doi.org/10.1002/acp.2890

Hoogerheide, V., Renkl, A., Fiorella, L., Paas, F., & van Gog, T. (2019). Enhancing example–based learning: Teaching on video increases arousal and improves problem–solving performance. *Journal of Educational Psychology, 111*, 45-56. https://doi.org/10.1037/edu0000272

Hoogerheide, V., Vink, M., Finn, B., Raes, A., & Paas, F. (2018). How to bring the news… Peak–end effects in children's affective responses to peer assessments of their social behavior. *Cognition & Emotion, 32*, 1114-1121. https://doi.org/10.1080/02699931.2017.1362375

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: adding a better end. *Psychological Science, 4*, 401-405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: A comparison of different objective measures. *Instructional Science, 45*, 515-536. https://doi.org/10.1007/s11251-017-9413-5

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 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*, 416-428. 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. https://doi.org/10.3758/s13421-017-0771-7

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

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

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*, 36-69. https://doi.org/10.1037/0096-3445.135.1.36

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

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80,* 325–340. https://doi.org/10.1348/000709910X485719

Larmuseau, C., Vanneste, P., Cornelis, J., Desmet, P., & Depaepe, F. (2019). Combining physiological data and subjective measurements to investigate cognitive load during complex learning. *Frontline Learning Research*, *7*, 57-74. https://doi.org/10.14786/flr.v7i2.403

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 464–470. https://doi.org/10.1037/0278-7393.16.3.464

Leppink, J., Paas, F., van der Vleuten, C. P., van Gog, T., & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*, 1058-1072.

Lord, F. M., & Novick, M. R.(1968): *Statistical theories of mental test scores*. Reading, Mass. USA: Addison-Wesley. https://doi.org10.3758/s13428-013-0334-1

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174-179. https://doi.org/10.3758/PBR.15.1.174

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

Mirza, F., Agostinho, S., Tindall-Ford, S., Paas, F., & Chandler, P. (2019). Self-management of cognitive load. In S. Tindall-Ford, S. Agostinho, & J. Sweller (Eds.), *Advances in Cognitive Load Theory: Rethinking Teaching* (pp. 157-167). London, UK: Routledge.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors*, *45*, 575-590. https://doi.org/10.1518/hfes.45.4.575.27094

Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems, 7*, 1-20. https://doi.org/10.1145/2960413

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*, 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*, 63-71. https://doi.org/10.1207/S15326985EP3801_8

Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G., & Darabi, A. A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner

involvement in instruction. *Educational Technology Research and Development*, *53*, 25-34. https://doi.org/10.1007/BF02504795

Paas, F., & van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Research*, *6*, 351–371. https://doi.org/10.1007/BF02213420

Paas, F., & Van Merriënboer, J. J. G. (1994b). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133. https://doi.org/10.1037/0022-0663.86.1.122

Paas, F., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*, 419-430. https://doi.org/10.2466/pms.1994.79.1.419

Peng, Y., & Tullis, J. G. (in press). Theories of intelligence influence self-regulated study choices and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0000740

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron, 72*, 692–697. https://doi.org/10.1016/j.neuron.2011.11.001

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

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

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology*, *37*, 688-695. https://doi.org/10.1093/geronj/37.6.688

Richter, J., & Scheiter, K. (2019). Studying the expertise reversal of the multimedia signaling effect at a process level: Evidence from eye tracking. *Instructional Science*, *47*, 627-658. https://doi.org/10.1007/s11251-019-09492-3

Richter, J., Scheiter, K., & Eitel, A. (2016). Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis. *Educational Research Review*, *17*, 19-36. https://doi.org/10.1016/j.edurev.2015.12.003

Richter, J., Scheiter, K., & Eitel, A. (2018). Signaling text–picture relations in multimedia learning: The influence of prior knowledge. *Journal of Educational Psychology*, *110*, 544–560. https://doi.org/10.1037/edu0000220

Rivers, M. L., Dunlosky, J., & Joynes, R. (2019). The contribution of classroom exams to formative evaluation of concept-level knowledge. *Contemporary Educational Psychology*, *59*, 101806. https://doi.org/10.1016/j.cedpsych.2019.101806

Ruiz, N., Taib, R., Shi, Y., Choi, E., & Chen, F. (2007). Using pen input features as indices of cognitive load. In *Proceedings of the ninth international conference on multimodal interfaces* (pp. 315–318).

Scharinger, C. (2018). Fixation-related EEG frequency band power analysis: A promising methodology for studying instructional design effects of multimedia learning material. *Frontline Learning Research, 6*, 57-71. https://doi.org/10.14786/flr.v6i3.373

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015). Pupil dilation and EEG alpha frequency band power reveal load on executive functions for link-selection processes during text reading. *Plos One, 10*, e0130608. https://doi.org/10.1371/journal.pone.0130608

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*, 93-114. https://doi.org/10.1007/s11251-014-9328-3

Schnotz, W., Fries, S., & Horz, H. (2009). Motivational aspects of cognitive load theory. In M. Wosnitza, S. A. Karabenick, A. Efklides, & P. Nenniger (Eds.), *Contemporary motivation research: From global to local perspectives* (p. 69–96). Hogrefe & Huber Publishers.

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review*, *24*, 116-129. https://doi.org/10.1016/j.edurev.2018.03.004

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*, 127-190. https://doi.org/10.1037/0033-295X.84.2.127

Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61-73. https://doi.org/10.1016/j.learninstruc.2017.01.002

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, *7*, 218. https://doi.org/10.3389/fpsyg.2016.00218

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science, 29*, 549-571. https://doi.org/10.1177/0956797617739704

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 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, 261–292. https://doi.org/10.1007/s10648-019-09465-5

Szulewski, A., Kelton, D., & Howes, D. (2017). Pupillometry as a tool to study expertise in medicine. *Frontline Learning Research*, *5*, 53-63. https://doi.org/10.14786/flr.v5i3.256

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

Thompson, V., Prowse Turner, J., Pennycook, G., Ball, L., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. Cognition, 128, 237-251. https://doi.org/10.1016/j.cognition.2012.09.012

Topolinski, S., & Reber, R. (2010). Gaining insight into the "Aha" experience. *Current Directions in Psychological Science*, *19*, 402-405. https://doi.org/10.1177/0963721410388803

Tsang, P., & Wilson, G.F. (1997). Mental workload. In G. Salvendy (Hrsg.), *Handbook of human factors and ergonomics* (pp. 417 – 449). New York, NY: Wiley.

Undorf, M., Livneh, I., & Ackerman, R. (2020). Help seeking as a metacognitive strategy when answering knowledge questions. *Manuscript submitted for publication.*

van Gog, T., Hoogerheide, V., & van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review*.

van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*, 833-839. https://doi.org/10.1002/acp.2883

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*, 16-26. https://doi.org/10.1080/00461520701756248

van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction*, *16*, 154-164. https://doi.org/10.1016/j.learninstruc.2006.02.003

van Harsel, M., Hoogerheide, V., Verkoeijen, P., & van Gog, T. (2019). Effects of different sequences of examples and problems on motivation and learning. *Contemporary Educational Psychology, 58*, 260-275. https://doi.org/10.1016/j.cedpsych.2019.03.005

van Merriënboer, J. J. G., Schuurman, J. G., De Croock, M. B. M., & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction*, *12*, 11-37. https://doi.org/10.1016/S0959-4752(01)00020-2

Winne P. H., Hadwin A. F. (1998). Studying as self-regulated engagement in learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 277-304). Hillsdale, NJ: Erlbaum.

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & Stress*, *14*, 74-99. https://doi.org/10.1080/026783700417249

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, 145, 918–933. http://doi.org/10.1037/xge0000177

Zu, T., Hutson, J., Loschky, L. C., & Rebello, N. S. (2019). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000441