

**Reference:**

Hoch, E., Sidi, Y., Ackerman, R., Hoogerheide, V., & Scheiter, K. (in press). Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective. *Educational Psychology Review*.

**Comparing Mental Effort, Difficulty, and Confidence Appraisals  
in Problem-Solving: A Metacognitive Perspective**

Emely Hoch<sup>\*1</sup>, Yael Sidi<sup>\*2</sup>, Rakefet Ackerman<sup>3</sup>, Vincent Hoogerheide<sup>4</sup>, and Katharina Scheiter<sup>1,5</sup>

\* First authorship is shared. Both authors contributed equally to this work.

<sup>1</sup> Leibniz-Institut für Wissensmedien, Tübingen, Germany


<sup>2</sup> Department of Education and Psychology, The Open University of Israel, Ra'anana, Israel

<sup>3</sup> Technion – Israel Institute of Technology, Haifa, Israel


<sup>4</sup> Department of Education, Utrecht University, Utrecht, the Netherlands


<sup>5</sup> Educational Science Department, University of Potsdam


**Author Note**

Emely Hoch  <https://orcid.org/0000-0002-6534-1506>

Yael Sidi  <https://orcid.org/0000-0003-2503-9166>

Rakefet Ackerman  <https://orcid.org/0000-0001-9583-8014>

Vincent Hoogerheide  <https://orcid.org/0000-0003-4176-0973>

Katharina Scheiter  <https://orcid.org/0000-0002-9397-7544>

Correspondence concerning this article should be addressed to Emely Hoch, Leibniz-Institut für Wissensmedien, Schleichstraße 6, 72076 Tübingen, Germany; phone: +49 7071 979-353; Email: e.hoch@iwm-tuebingen.de.

### **Acknowledgments**

This research was supported and inspired by discussions of the EARLI Emerging Field Group Monitoring and Regulation of Effort.

### **Statements and Declarations**

#### **Funding**

This research was funded by the Jacobs Foundation in the context of the EARLI Emerging Field Group Monitoring and Regulation of Effort and by the Israel Science Foundation.

#### **Conflicts of Interest/Competing Interests**

Rakefet Ackerman is an editorial board member of Educational Psychology Review. Otherwise, the authors have no competing interests to declare relevant to this article's content.

#### **Ethics Approval**

The ethical conduct of the studies was reviewed and approved by the Behavioral Sciences Research Ethics Committee of the Technion – Israel Institute of Technology (2020-015) and the Leibniz-Institut für Wissensmedien Institutional Review Board (protocol LEK 2021/011).

#### **Availability of Data**

The datasets generated and/or analyzed during the current study along with the corresponding analysis syntax are available in the OSF repository, [https://osf.io/q2p74/?view\\_only=095cdcf16a314cbd8334ad249ba6fc70](https://osf.io/q2p74/?view_only=095cdcf16a314cbd8334ad249ba6fc70).

#### **Authors' Contributions**

All authors contributed to the studies' conception and design. Material preparation, data collection, and analysis were mainly performed by Yael Sidi (Experiment 1), Rakefet Ackerman (Experiment 2), and Emely Hoch (Experiment 3). The first draft of the manuscript was written by Emely Hoch and Yael Sidi and commented on by all authors. All authors read and approved the final manuscript.

#### **Consent to Participate**

Informed consent was obtained from all individual participants included in the study.

### **Abstract**

It is well-established in educational research that metacognitive monitoring of performance assessed by self-reports—for instance, asking students to report their confidence in provided answers—is based on heuristic cues rather than on actual success in the task. Subjective self-reports are also used in educational research on cognitive load, where they refer to the perceived amount of mental effort invested in or difficulty of each task item. In the present study, we examined the potential underlying bases and the predictive value of mental effort and difficulty appraisals compared to confidence appraisals by applying metacognitive concepts and paradigms. In three experiments, participants faced verbal logic problems or one of two non-verbal reasoning tasks. In a between-participants design, each task item was followed by either mental effort, difficulty, or confidence appraisals. We examined the associations between the various appraisals, response time, and success rates. Consistently across all experiments, we found that mental effort and difficulty appraisals were associated more strongly than confidence with response time. Further, while all appraisals were highly predictive of solving success, the strength of this association was stronger for difficulty and confidence appraisals (which were similar) than for mental effort appraisals. We conclude that mental effort and difficulty appraisals are prone to misleading cues like other metacognitive judgments and are based on unique underlying processes. These findings challenge the accepted notion that mental effort appraisals can serve as reliable reflections of cognitive load.

*Keywords:* self-regulated learning, metacognition, cognitive load, effort, monitoring, meta-reasoning

Educational research on learning and instruction has developed along two distinct paths that, until recently, have acted largely in isolation (de Bruin et al., 2020; de Bruin & van Merriënboer, 2017). First, research on *self-regulated learning* (SRL) has focused on how students plan for a learning task, monitor their performance, and then reflect on the outcome, either spontaneously or with guidance (e.g., Zimmerman, 2002). Within this domain, *metacognitive research* (see Fiedler et al., 2019, for a review; Nelson & Narens, 1990) aims to expose conditions under which self-appraisals of knowledge may be biased, and the consequences of such bias for subsequent learning-regulation decisions (e.g., allocation of study time, use of study strategies, and help-seeking). Second, instructional design research focuses on developing learning tasks that support effective knowledge acquisition (e.g., Richter et al., 2016; van Gog, 2022). Much of this research has been conducted against the backdrop of *cognitive load theory* (CLT; Chandler & Sweller, 1991), which focuses on optimizing effort investment in learning or task performance.

The evident potential of these two massive bodies of research to fertilize each other has drawn attention in recent years (Baars et al., 2020; Blissett et al., 2018; de Bruin et al., 2020; Scheiter et al., 2020; Seufert, 2020; van Gog et al., 2020). In a recent review, Scheiter et al. (2020) highlighted that metacognitive and CLT research share the use of subjective self-appraisals of the learning process and learning outcomes. In particular, in metacognitive research, participants are asked to rate (or predict) their own expected or perceived performance immediately before or after performing each task item (e.g., solving a problem). These ratings take the form of metacognitive judgments, such as ease of learning, judgments of learning, feeling of rightness, or confidence (see Ackerman & Thompson, 2015). For consistency with CLT terminology, hereafter we refer to these judgments as *appraisals*. Metacognitive research has systematically shown that such appraisals are prone to biases, as they are based on heuristic cues and lay theories (Ackerman, 2019; Koriat et al., 2008). A massive body of metamemory and meta-reasoning research has exposed how inferential cues (mis)guide metacognitive monitoring (e.g., Ackerman & Beller, 2017; Bjork et al., 2013; Castel, 2008; Koriat, 2008; Undorf, 2020). It is also well-established that metacognitive appraisals guide (and thus may mislead) self-regulation decisions (for a review, see Fiedler et al., 2019). Similarly, in classic studies based on CLT, participants were asked to report the amount of effort they invested and/or the difficulty they experienced as indicators of the cognitive load associated with a particular task design (i.e., load-related appraisals). These self-appraisals are taken as indicators of the effectiveness of the given instructional design.

It seems reasonable to consider load-related appraisals as a type of metacognitive judgment (see Scheiter et al., 2020). From a metacognitive perspective, CLT appraisals, like other documented metacognitive appraisals, are presumably based on heuristic cues, and thus are prone to biases (Ackerman, 2019; Koriat, 1997). In the present study, we examine the processes that underlie load-related appraisals, with the goal of exposing whether they are prone to bias in the same way as known metacognitive appraisals, as well as their predictive value for task outcomes. Towards this end, we utilized metacognitive concepts and research methodologies.

### **Metacognitive Appraisals**

Metacognitive processes accompany the full course of cognitive activities involved in self-regulated learning, taking place spontaneously in parallel to knowledge processing. Researchers distinguish between two types of metacognitive processes: metacognitive monitoring and metacognitive control. Metacognitive monitoring refers to activities aimed at tracking, reviewing, and assessing the quality of one's cognition, while metacognitive control refers to decision-making about actions to be taken based on the outputs of those monitoring operations (for a review, see Fiedler et al., 2019; Nelson & Narens, 1990). For example, when solving a mathematical problem, one assesses the likely correctness of the solution that comes to mind. Based on this assessment, the solver decides whether to provide this solution or to invest more effort in searching for another solution (Efklides, 2008). An unreliable assessment impairs the consequent decision. Thus, to be effective, control decisions must be based on reliable monitoring (Ackerman & Thompson, 2017). In empirical research, monitoring reliability is commonly examined by collecting subjective performance appraisals (e.g., confidence ratings for a provided solution), comparing them to objective performance measures (e.g., the participant's performance in the task), and measuring the correspondence between the two measures.

An essential theoretical framework for understanding monitoring processes and factors influencing their accuracy is the cue utilization approach, which originated in metacognitive research focused on memorization tasks (Koriat, 1997). This framework suggests that people do not objectively know their knowledge level for a given task or item but infer it from a complex set of heuristic cues. Koriat (1997) classified these heuristic cues as either intrinsic cues inherent to the study items (e.g., ease of processing, familiarity of items, concreteness of items) or extrinsic cues related to the learning context (e.g., number of times items were presented for study). Cue utilization is the extent to which each heuristic

cue is considered when making metacognitive appraisals. Metamemory research has extensively investigated this framework, demonstrating how cues guide metacognitive monitoring both uniquely and simultaneously (Ackerman & Beller, 2017; Castel, 2008; Koriat, 2008; Undorf et al., 2018). The cue utilization approach has also been extended to the domain of problem-solving (for a review and classification, see Ackerman, 2019; e.g., Finn & Tauber, 2015; Metcalfe & Finn, 2008a, 2008b; Sidi et al., 2017). In the realm of load-related appraisals, findings suggest that monitoring of effort is an inference-based process similar to metacognitive appraisals (e.g., Dunn, Gaspar, et al., 2019; Dunn & Risko, 2016; Koriat, 1997; Raaijmakers et al., 2017).

Notably, while some cues have been found to predict performance and effort reliably, others have been implicated in biasing the monitoring process. For example, one of the most prominent cues within the metacognitive literature is *answer fluency* (cf. processing fluency, Ackerman, 2019; Thompson, Turner, et al., 2013). Answer fluency reflects the ease of processing and relates to the momentary experience of ease or difficulty one feels while performing each task item (Ackerman, 2019). Answer fluency has been primarily studied in the context of memorization tasks (for a review, see Schwartz & Jemstedt, 2021), and more recently with reasoning and problem-solving tasks (e.g., Ackerman & Beller, 2017; Wang & Thompson, 2019). Answer fluency is often operationalized by measuring response time: i.e., how much time was invested in solving a particular item (question or problem). Overall, response time has been identified as a valid cue, showing inverse relationships with both performance and metacognitive appraisals in various cognitive tasks (Benjamin & Bjork, 1996; Hertwig et al., 2008). However, under some conditions, response time has been found to bias metacognitive appraisals (see Finn & Tauber, 2015 for a review). For example, Kelley and Lindsay (1993) primed participants with a list of words and then asked them to answer a general knowledge test. They found that the relationship between response time and confidence was similar—specifically, that participants assumed quickly retrieved answers were correct—for questions that were and were not constructed to be misleading (by including in the initial list a word that was related to the question yet was not the correct answer). Benjamin et al. (1998) showed that people rely on retrieval fluency, operationalized by response time, in a general knowledge task, even though response time did not correspond to their actual recall. This resulted in a negative relationship between appraisals and recall performance. In the domain of problem-solving, Ackerman and Zalmanov (2012) found that the association between solving time and confidence remained persistent even

when the speed at which problems were solved did not predict accuracy. All these findings suggest that relying on response time can misguide the monitoring process.

Turning to difficulty appraisals, Kelley and Jacoby (1996) showed that response time serves as a potentially misleading cue here as well. In their study, participants were asked to solve anagrams and, in some conditions, were pre-exposed to some of the solution words, resulting in shorter response times for these items. The correlation between response time and difficulty appraisals was consistently high across items and conditions. Kelley and Jacoby attributed this to the biased subjective experience of difficulty cued by response time. These findings raise the question, does response time have differential relationships with confidence, difficulty, and mental effort?

### **Inferring Cues for Mental Effort Appraisals From Metacognitive Research**

Cognitive load is “a multidimensional construct that represents the load that performing a particular task imposes on the cognitive system of a particular learner” (Paas & Van Merriënboer, 1994, p. 122). CLT’s central premise is that the capacity of human working memory to process novel information is limited. Therefore, instructional tasks should be designed to reduce unnecessary load and promote schema acquisition, organization, and automation (van Merriënboer & Kirschner, 2017). Although some educational research uses objective measures of cognitive load (e.g., Chen et al., 2022; Korbach et al., 2018; Szulewski et al., 2017), most such research has measured cognitive load via self-report measures of effort or difficulty (Naismith et al., 2015), as these are easier both to administer and to interpret.

Self-report cognitive load appraisals are meant to reflect the amount of capacity or resources allocated to accommodate the task demands (Brünken et al., 2003, 2006). CLT research has traditionally used effort investment and task difficulty interchangeably for this purpose (de Jong, 2010). Both types of appraisals are commonly elicited using 5-, 7-, or 9-point Likert scales (with the 9-point Likert scale being the one initially proposed by Paas, 1992). Task difficulty scales typically use wording like (e.g., “the task was very, very easy ... very, very difficult”; e.g., Ayres, 2006). Effort appraisals have two common variations. One focuses on the person’s voluntary investment of effort (e.g., “I invested very, very low mental effort ... very, very high mental effort”; e.g. Paas, 1992; cf. van Gog & Paas, 2008),

and the other focuses on the task as requiring low/high effort (e.g., “The task required very, very low... very, very high effort”).<sup>1</sup>

Scheiter et al. (2020) conceptualized the various CLT appraisal items from a metacognitive perspective, arguing that the phrasing of effort investment items can reflect motivational and cognitive aspects related to processing and task performance. Particularly, *the effort the individual decided to invest* can be referred to as goal-driven effort (Koriat et al., 2006). Goal-driven effort relies on top-down processing, reflecting voluntary decisions made by learners. In contrast, *the effort the task demands* can be referred to as data-driven effort (Koriat et al., 2006). It is based on bottom-up processing, focusing on task characteristics that learners cannot control, similar to asking about the difficulty of the task.

Self-appraisals of effort and difficulty are utilized in CLT research under the assumption that they reliably reflect the cognitive resources people allocate to the task. Yet evidence suggests that load-related appraisals might partly reflect biases that stem from unreliable cues (see Scheiter et al., 2020, for a review). For instance, Raaijmakers et al. (2017) examined the effects of performance feedback as an external heuristic cue for mental effort appraisals in a complex problem-solving task. In their study, half the participants received positive feedback, and the other half received negative feedback, irrespective of actual task performance. In three experiments, feedback indeed affected effort appraisals, with the direction depending on feedback valence: positive feedback (pointing to success) was related to lower effort appraisals than negative feedback (pointing to failure). Other studies focused on the timing and frequency of load-related appraisals (Ashburner & Risko, 2021; Schmeck et al., 2015; van Gog et al., 2012). In one study (van Gog et al., 2012), single delayed appraisals provided at the end of a series of tasks yielded higher cognitive load estimates than the average of appraisals provided immediately after each task item, regardless of performance. This effect was particularly pronounced for the more complex tasks in the set. Relatedly, Ashburner and Risko (2021) demonstrated that post-trial appraisals were associated with perceptions of greater effort compared to post-whole task appraisals, regardless of objective task demands. Looking at cues inherent to the task, Dunn

---

<sup>1</sup> CLT also distinguishes between three types of cognitive load (intrinsic load, extraneous load, and germane load), for each of which recent research has developed and validated unique measures (e.g., Klepsch et al., 2017; Klepsch & Seufert, 2020; Leppink et al., 2013). However, investigating the different types of cognitive load was not the focus of this study. Moreover, the distinction itself, as well as the construct of germane load as a distinguishable category of cognitive load, are under debate in the CLT literature (Sweller et al., 2019). As the present study is an initial investigation of the bases of self-reported effort and difficulty, we focus on the classic framing of mental effort, and rely on the most common measures used in CLT research.



and Risko (2016) examined effort appraisals for reading in four types of display conditions, involving rotations to either the presented words, the frame, neither, or both. Their findings showed that participants evaluated displays in which both the words and frame were rotated as more effortful to process than those in which only the words were rotated. However, these appraisals were dissociated from actual performance, as in fact the real difference was between all displays in which the words were rotated and those where only the frame was rotated, with performance being better in the latter. These findings demonstrate that load-related appraisals may rely on external cues, expressing sensitivity to task demands, while being dissociated from objective measures of success.

Might response time be another culprit biasing load-related appraisals, similar to how response time has been found to bias performance appraisals (e.g., confidence, Finn & Tauber, 2015)? While this question has not been directly investigated, related research does suggest such a link (e.g., Leppink & Pérez-Fuster, 2019). For example, Dunn et al. (2019) drew an association between perceived task effort and task time requirements. They compared effort appraisals (how “effortful” the task is) when a decision-making task presented participants with competing “costs”—the time required by a task (low or high), and how error-prone the task was. While error likelihood was more strongly associated with effort appraisals, Dunn et al. reported that both costs predicted effort appraisals across several experimental conditions.

Task complexity has also been investigated in CLT research in relation to cognitive load. In particular, tasks involving more interacting elements that need to be stored in working memory impose a higher load (Sweller et al., 1998). Taking a metacognitive perspective raises the question of whether load-related appraisals are guided by task complexity. Specifically, do people acknowledge the effect of variations in task complexity on the load it imposes on their working memory? Haji et al. (2015) investigated the sensitivity of goal-driven effort appraisals and response time to task complexity in simulation-based surgical skills training by comparing two groups faced with low and high complexity levels. Their low-complexity group provided lower effort appraisals than the high-complexity group, with no corresponding differences in response time. This serves as an initial indication that complexity could serve as a cue for load-related appraisals. However, it is still unclear whether complexity also serves as a cue for load appraisals when complexity varies between different items within a single task. Also, notably, research has suggested that the association between item-level complexity and response time is not

straightforward, as people may not be motivated to invest the required effort for solving highly complex task items (e.g., Ackerman, 2014; Hawkins & Heathcote, 2021; Paas et al., 2005).

Taken together, these findings expose the need to systematically investigate how response time and task (or item-level) complexity are associated with the different mental effort appraisals, in order to infer their strength as heuristic cues for mental effort appraisals compared to metacognitive appraisals.

### **The Predictive Value of Effort and Difficulty Appraisals**

Monitoring accuracy has been a central factor in SRL theory and research due to its causal role in guiding subsequent metacognitive control decisions (Panadero, 2017; Winne & Perry, 2000). While the effectiveness of SRL is strongly dependent on the predictive value of load-related appraisals, this has yet to be systematically examined (de Bruin et al., 2020). As indicated above, research has shown initial evidence that load-related appraisals rely on contextual factors other than the mental effort involved or the difficulty of the task (Raaijmakers et al., 2017; Schmeck et al., 2015; van Gog et al., 2012). For example, Rop et al. (2018) showed that mental effort ratings decreased with increasing task experience; however, the results regarding success in the task were inconsistent across two experiments. This finding suggests that the alignment between effort ratings and performance may change over time. In the present study, we aimed to delve into the predictive value of load-related appraisals, namely their association with task success.

As Scheiter et al. (2020) suggested, one way to consider the predictive value of load-related appraisals is by looking into the predictive power of an established external criterion. In metacognitive research, the criterion used to validate subjective task appraisals—monitoring accuracy—is the success rate in performing the task at hand. This is done using two measures: calibration and resolution. *Calibration* represents the overall fit between subjective appraisals and actual performance. It can be biased either upwards, resulting in overconfidence, or downwards, resulting in underconfidence. Calibration bias can mislead effort regulation and result in inferior learning outcomes (Metcalfe & Finn, 2008b; Thiede et al., 2003). *Resolution* represents the extent to which people distinguish in their confidence between correct and incorrect responses. It is measured at the individual level as a within-participant correlation between confidence and success in each item. Scheiter et al. (2020) maintained that of the two measures, resolution is the more relevant for load-related appraisals due to it being a relative measure (referring to the variability of ratings across task

items) rather than an absolute measure (using the numerical value of each rating). Thus, in the present study, we examined the resolution of the various load-related appraisals compared to metacognitive confidence appraisals.

### **Research Questions and Study Overview**

Following this review, we aimed to examine three main research questions.

*RQ1. Are there differences in the extent to which response time serves as a cue for mental effort appraisals, difficulty appraisals, and confidence appraisals?*

The literature reviewed above suggests that all these types of appraisals are guided to some extent by response time. However, no research thus far has compared these relationships within one study to examine their relative strength. As fluency, by definition, relates to the momentary experience of ease or difficulty, we expected that all load-related appraisals would be found to rely more strongly on response time as a cue compared to confidence. This is supported by evidence in the metacognitive literature that confidence shows only a modest relationship with response time (Ackerman, 2014).

*RQ2. Are there differences in the extent to which mental effort appraisals, difficulty appraisals, and confidence appraisals predict actual accuracy in each task item?*

We expected that confidence would predict task accuracy, based on ample prior research. How the different mental effort appraisals relate to task accuracy is less clear. Yet this is important to examine, because there are reasons to believe that goal-driven effort and data-driven effort may show unique relationships with task accuracy. As explained above, Koriat et al.'s (2006) theory suggests that goal-driven effort, operationalized by additional time invested when motivation to succeed rises, reflects the voluntary decision to invest resources into a task. This could result in investing effort in vain due to higher internal motivation to succeed, which might not necessarily result in actual higher success. However, data-driven effort, according to Koriat et al.'s theory, reflects task demands in a similar manner to confidence. Therefore, one could expect data-driven effort to better predict task accuracy compared to goal-driven effort.

The association of perceived difficulty with task accuracy largely relies on how people interpret requests for difficulty appraisals. Do they believe they are being asked about the amount of effort they felt they personally had to invest, or about difficulty as a facet of the problem? This is an open question on which our study can shed light by comparing the strength of the relationship between difficulty to task accuracy and between difficulty and response time with relations of different appraisals.

*RQ 3: Are there differences in the extent to which the complexity of the problem serves as a cue for mental effort appraisals, difficulty appraisals, and confidence appraisals?*

Empirical evidence from both CLT research and metacognitive research has shown that both load-related appraisals and confidence are sensitive to differences in task elements related to complexity (e.g., Ayres, 2006; Paas & Van Merriënboer, 1994; Schmeck et al., 2015; van Gog et al., 2012). However, no studies have yet compared load-related and confidence appraisals in terms of their sensitivity to complexity, leaving this an open research question.

This study was designed to test these research questions systematically, using three experiments. All three experiments were designed to examine RQ1, on response time as a cue for the different appraisals, and RQ2, on the predictive value of the various appraisals for success in the task. Experiment 3 also examines RQ3, on item complexity. In all three experiments, participants completed reasoning and problem-solving tasks in a multiple-choice test format, followed by mental effort, task difficulty, or metacognitive confidence appraisals. While the tasks differed between the experiments, the designs and procedures were very similar. All experiments included four groups differing only in the type of appraisal provided immediately after completing each task item: goal-driven effort, data-driven effort, task difficulty, or metacognitive confidence. To examine our research questions, we analyzed the associations between the various appraisals and the relevant outcome of interest (response time, task accuracy, or item complexity), as appropriate. In the first experiment, we employed a verbal logic task widely used in cognitive and metacognitive research—the Cognitive Reflection Test (CRT; Frederick, 2005). The task consists of misleading verbal mathematical problems (word problems) designed so that the first solution that usually comes to mind is an incorrect but predictable one, while most respondents can arrive at the correct solution with more effort investment. The task calls for heterogeneous appraisals across items, which is essential for examining within-participant correlations between appraisals on the one hand, and response times or accuracy on the other.

Notably, most empirical research on meta-memory processes has used verbal tasks, like memorizing words and answering knowledge questions. When considering meta-reasoning tasks, verbal tasks dominate research as well, with the CRT, as used in Experiment 1, being an example (see Ackerman & Thompson, 2017, for examples and a review). The

scarce studies utilizing non-verbal tasks show some similarities in the metacognitive mechanisms involved (Lauterman & Ackerman, 2019; Reber et al., 2008). To contribute to studying non-verbal reasoning processes and examine our findings' robustness, in Experiment 2, we used a non-verbal problem-solving task: the Missing Tan Task (MTT; Ackerman, in press). The MTT is a challenging non-verbal reasoning task that relies on cognitive processes also involved in geometry, navigation, and design. Participants are presented with silhouettes generated from geometric pieces (called tans). The silhouettes are comprised of six pieces, drawn from a pool of seven. Participants' task is to identify which piece is not needed to form each silhouette solely through mental visualization, without being able to manipulate the presented pieces.

In the third experiment, we replicated Experiment 1 and Experiment 2 using yet another non-verbal task. In order to also address RQ3, for this experiment we chose a task which offers inherent variations in item complexity: the mental rotation task (MRT; Shepard & Metzler, 1971). Here, participants are presented with a set of rotated stimuli, and must mentally rotate each one to align with a criterion stimulus in order to determine which alternative matches the criterion figure (Searle & Hamm, 2017). Based on previous research, we used the angle of rotation as an objective measure of item complexity (Shepard & Metzler, 1971). Complexity from a CLT perspective usually refers to the number of interacting elements of a task that are processed simultaneously in working memory; the greater the number of interacting elements, the higher the cognitive load (e.g., van Gog & Sweller, 2015). In the present task, we assume that the larger the rotation angle (i.e., the more mental rotation required), the higher the cognitive load.

While the three tasks rely on different reasoning skills, they are similar in several ways, supporting comparisons between the findings. (a) The tasks call for deliberate reasoning processes. (b) The tasks are in a multiple-choice format. (c) The tasks allow for the generation of a wide variety of items, resulting in different success rates, while the time spent processing the stimuli remains similar across items. This feature is important for the variability in response time to reflect its variability within participants across success rates. (d) The tasks involve uncertainty as to whether the solution provided is correct (unlike, e.g., fitting a jigsaw puzzle piece into the right spot, which usually involves no uncertainty). (e) Having more than fifteen items allows for robust within-participant statistical analyses.

Finally, in all experiments, we collected data on individual differences in participants' self-perceptions or beliefs about their traits, abilities, or knowledge. More

specifically, (a) in Experiment 1 participants reported on their need for cognition (Cacioppo & Petty, 1982). This scale reflects cognitive style, or the extent to which the individual enjoys taking part in effortful cognitive activities. In Experiments 2 and 3 participants completed scales capturing (b) test anxiety (Taylor & Deane, 2002), reflecting self-doubt about their ability to succeed in a particular task type, and (c) beliefs about the malleability of intelligence (Dweck et al., 1995), reflecting implicit theories of intelligence as malleable (growth mindset) or unmalleable (fixed mindset). Metacognitive research has shown that these constructs are associated with confidence or similar appraisals of expected success (e.g., Jonsson & Allwood, 2003; Kirk-Johnson et al., 2019; Miele et al., 2011; Miesner & Maki, 2007; Petty et al., 2009). Thus, we sought to examine them as potential moderators for how the different appraisals are related to response time and accuracy (Scheiter et al., 2020).

### **Experiment 1**

Experiment 1 aimed to examine the associations between the various appraisals and response time, as well as with task accuracy, by applying a verbal logic task.

#### **Method**

##### ***Participants and Design***

Data were collected online through the Prolific ([www.prolific.co](http://www.prolific.co)) participant pool. Participants were required to be at least 20 years old, to speak English fluently (to ensure they understood the instructions), and to have no learning disabilities. Participation was voluntary, anonymous, and remunerated with 2GBP. We excluded data from participants who encountered technical problems during the experiment (6 participants); did not follow instructions (e.g., admitted to having engaged in other activities while completing the tasks; 5 participants); had little variability in appraisals (i.e.,  $SD < 4$ ; 4 participants); or provided valid responses to less than 75% of the trials (see exclusion criteria for single trials under Data Preparation; 7 participants). The final sample comprised data from 284 participants (age:  $M = 34.5$  years,  $SD = 10.9$ ; 146 females, 125 males; age and gender missing for 13 participants). Participants were randomly assigned to one of four groups that differed in the type of appraisal they were asked to provide: goal-driven effort ( $n = 70$ ), data-driven effort ( $n = 71$ ), task difficulty ( $n = 70$ ), and confidence ( $n = 73$ ).

##### ***Materials: Misleading Math and Logic Problems (CRT Tasks)***

The original CRT (Frederick, 2005) contains three misleading math problems, where the first solution that commonly comes to mind is a wrong but predictable one, but a little deliberative effort can lead most respondents to the correct solution. While the CRT is

suitable for examining our research question, the original CRT problem set is so widely used as to raise concerns regarding participants' pre-exposure to the task. Also, to allow robust within-participant statistical analyses, it was essential to have a larger number of task items. Therefore, for the present study we used a collection of 17 misleading math and logic problems based on several resources, fitted to a multiple-choice format, and pretested (see Appendix A). For instance, one of the items was "25 soldiers are standing in a row 3 m from each other. How long is the row?" Participants had to choose from four answers: a) 3 m, b) 69 m, c) 72 m, d) 75 m. The answer, which is expected to jump quickly to mind, 75 m, is wrong. The correct answer is 72 m (Oldrati et al., 2016).

### ***Appraisals***

Each participant provided one of four appraisals for all items in the study, depending on their experimental group: goal-driven effort, data-driven effort, task difficulty, or confidence. Participants entered their appraisal by sliding a bar on a horizontal slider using the mouse. All scales ranged from 0 to 100. In the goal-driven effort condition the question was "How much effort did you invest in solving the problem?", and the scale ranged from *very, very low effort* (0) to *very, very high effort* (100). In the data-driven effort condition the scale anchors were the same, but participants were asked, "How much effort did the problem require?" In the task difficulty condition the question was "How difficult was the problem?", and the scale anchors were *very, very easy* (0) and *very, very difficult* (100). Finally, participants in the confidence condition were asked, "How confident are you that your solution is correct?", on a scale from *a wild guess* (0) to *definitely sure* (100). The scales were adapted from metacognitive research, in which such scales are commonly used for different types of metacognitive appraisals. Their advantage is their receptiveness to comparison with actual task performance (also ranging from 0 to 100), which allows calculating monitoring accuracy.

### ***Objective Measures***

Response time was defined as the time each participant took to respond to each task item (in seconds). Our second objective measure was item-level task accuracy or providing a correct answer for each task item.

### ***Background Variables<sup>2</sup>***

Previous knowledge<sup>3</sup> about the task was assessed by this question: “Have you ever encountered one or more of the problems that you solved here in other studies? If you did, please write what you remember from those problems. If not, please enter ‘All new.’” Need for cognition (Cacioppo & Petty, 1982; Lins de Holanda Coelho et al., 2020) was measured with the short (six-item) scale assessing the extent to which people enjoy engaging in the process of thinking (e.g., “I really enjoy a task that involves coming up with new solutions to problems”). Responses were given on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*), with two items reverse-coded (Cronbach’s  $\alpha = .84$ ). Experience with solving puzzles (“How often do you solve puzzles or play thought-provoking games?”) was assessed with a single item on a scale from 1 (*almost never*) to 7 (*daily*).

One item serving as an attention check was presented amidst the need for cognition questions, using the same scale (“I like taking logic exams. Please ignore this statement, wait at least four seconds, and then respond by level two”). Less than 6% of participants failed to pass the attention check. However, these participants were only excluded if they also showed another indication of inattention (failure to follow study instructions; see under Participants and Design).

All means and standard deviations of the background variables as a function of the type of appraisal are shown in Appendix B.

### ***Procedure***

Participants first received general information about the procedure and gave their consent to participate. They then received the instructions for the CRT task—to solve verbally phrased math and logic problems by choosing one out of four solution options. In the training phase, participants solved an example problem for which the correct answer was provided. With a second example, one of the four appraisals corresponding to the relevant condition (goal-driven effort, data-driven effort, task difficulty, confidence) was introduced. Participants were told that as the problems were not trivial, ratings across the entire range of the scale, including low and intermediate rating levels, were expected. Following the examples, participants solved the 17 CRT items in random order. Work on the items was

---

<sup>2</sup> Background variables were also exploratively examined as possible moderators in all three experiments. However, since few significant results and, in particular, no consistent patterns emerged, the results are not reported for the sake of brevity.

<sup>3</sup> Four participants stated that they had already encountered most or all of the items. However, since excluding them from the analyses did not change the pattern of results, these participants were kept in the sample to maintain statistical power.



self-paced. Once a solution option was selected, it could not be changed, and an appraisal was elicited. After completing all CRT items, participants answered the background questions. Then, participants were asked whether they had pursued other activities during the experiment, had additional comments, or had encountered technical problems. Participants completed the experiment in about 15 minutes on average.

### ***Data Preparation***

Data preparation and all analyses were conducted in R (R Core Team, 2021). Trials with extraordinarily short ( $RT < 2$  sec; 11 trials) or extraordinarily long response times ( $RT > 180$  sec; 130 trials) were excluded, as were trials where participants left the experimental environment to do something else (168 trials). This left 4,752 trials included in the data analyses.

As we were interested in the strength of the relationship between different appraisals and objective measures, this association was calculated as the within-participant correlation across items between each appraisal and the objective measures of response time and accuracy. As the objective measures under investigation differed in their scale levels, Pearson correlations were used to calculate the correlation between response time (an interval scaled variable) and appraisals, while the Goodman-Kruskal  $\gamma$  rank correlation was used to calculate the correlations between task accuracy (a dichotomous variable) and appraisals. Note that differences between the appraisals were expected simply because of their opposing reference points: easy items should naturally yield high confidence values, but low values for task difficulty and effort. Thus, to statistically address RQ1 and RQ2, confidence appraisals were reversed to match the other appraisals' direction of association with item difficulty. It should also be noted that the correlations are interpreted differently for RQ1 and RQ2. In RQ1, the correlation indicates the degree to which response time serves as a cue for the appraisal. In RQ2, the correlation reflects the extent to which the appraisal has predictive value for task accuracy.

### **Results and Discussion**

Item success rates (i.e., the percentage of participants who correctly solved a given item) ranged from 26.6% to 81.9%. On average, participants needed 24.7 sec ( $SD = 12.5$ ) for each item, and they correctly solved 56.0% ( $SD = 21.7$ ) of the items. Means and standard deviations of appraisals, response time, and success as a function of appraisal are shown in Table 1. The four appraisal groups did not differ in either their response times,  $F(3, 280) = 2.12$ ,  $MSE = 154.1$ ,  $p = .098$ ,  $\eta^2 = .02$ , or in accuracy,  $F < 1$ .

**Table 1**

*Means and Standard Deviations of Response Time, Accuracy, and Appraisals as a Function of Type of Appraisal in Experiments 1–3*

	Goal-driven effort	Data-driven effort	Task difficulty	Confidence
<u>Experiment 1 (CRT)</u>				
Response time	27.33 (16.94)	25.29 (12.12)	23.82 (10.31)	22.31 (8.92)
Accuracy	53.66 (20.77)	56.22 (21.56)	58.58 (21.42)	55.61 (23.28)
Appraisal	36.64 (21.94)	35.12 (19.14)	29.91 (15.33)	80.72 (9.42)
<u>Experiment 2 (MTT)</u>				
Response time	26.46 (14.43)	26.31 (15.84)	27.16 (16.05)	24.22 (13.57)
Accuracy	42.59 (16.20)	41.52 (13.34)	45.81 (17.54)	42.26 (16.65)
Appraisal	55.28 (16.43)	53.60 (15.23)	55.74 (14.72)	67.66 (12.52)
<u>Experiment 3 (MRT)</u>				
Response time	21.93 (10.62)	23.68 (12.33)	21.06 (10.81)	21.01 (12.34)
Accuracy	77.23 (20.19)	69.75 (22.82)	72.63 (23.14)	72.87 (23.68)
Appraisal	47.06 (20.80)	42.95 (14.08)	42.15 (14.91)	81.44 (12.19)

*Note.* Response time gives the average time (in seconds) for answering an item. Accuracy is calculated as the percentage of correct answers. Appraisals were given on a 0–100 scale. In Experiments 2 and 3, confidence appraisals were given on a 20–100 scale to account for the probability of getting the item right just by guessing.

To describe the relationships between the appraisals and the variables of interest (response time and accuracy), Table 2 presents the mean within-participant correlations as a function of appraisal. Unsurprisingly, as mentioned above, the mathematical signs for the correlations with confidence were the inverse of those for the other appraisals. To test whether these relationships are meaningful, the mean correlations were tested against 0. Adjusted *p*-values were calculated using Bonferroni correction to account for multiple tests (i.e., four tests for the associations between the appraisals and response time, and again between the appraisals and accuracy). As seen in Table 2, all correlations were significant except for one (the correlation between goal-driven effort and accuracy). These findings indicate that response time does indeed serve as a cue for appraisals, and that appraisals do

predict actual accuracy in the task. Therefore, it is plausible to examine differences between the four types of appraisals in both cases.

***RQ 1: Are There Differences in the Extent to Which Response Time Serves as a Cue for Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals?***

To evaluate whether the different appraisals rely on response time in similar ways (see the CRT columns in Figure 1A), the strength of the appraisal–response time correlation was compared between the different appraisal types using ANOVA<sup>4</sup> (confidence appraisals were reversed for this comparison). There was a significant large effect of appraisal type,  $F(3, 280) = 16.93$ ,  $MSE = 0.09$ ,  $p < .001$ ,  $\eta^2 = .15$ . Post-hoc pairwise comparisons ( $t$ -tests with Bonferroni adjustment for  $p$ -values) showed that the correlation with response time was significantly weaker for confidence than for each of the other three appraisal types (goal-driven effort, data-driven effort, and task difficulty, all  $p_s < .001$ ). All other comparisons showed no significant differences, all  $p_s > .05$ . These findings present initial evidence that response time serves as a cue for load-related appraisals, that is, goal- and data-driven effort appraisals as well as task difficulty appraisals.

---

<sup>4</sup> The assumption of normally distributed data was violated in all three experiments. However, in such cases, the  $F$ -statistic in fixed effects models is still considered robust when group sizes are equal (Lunney, 1970; Schmider et al., 2010). Furthermore, the assumption of homogeneity of variances was violated when answering RQ1 in Experiment 2. Since data transformation did not resolve this issue and using non-parametric alternatives to ANOVA did not change the pattern of results, we report the results of ANOVA throughout the manuscript.

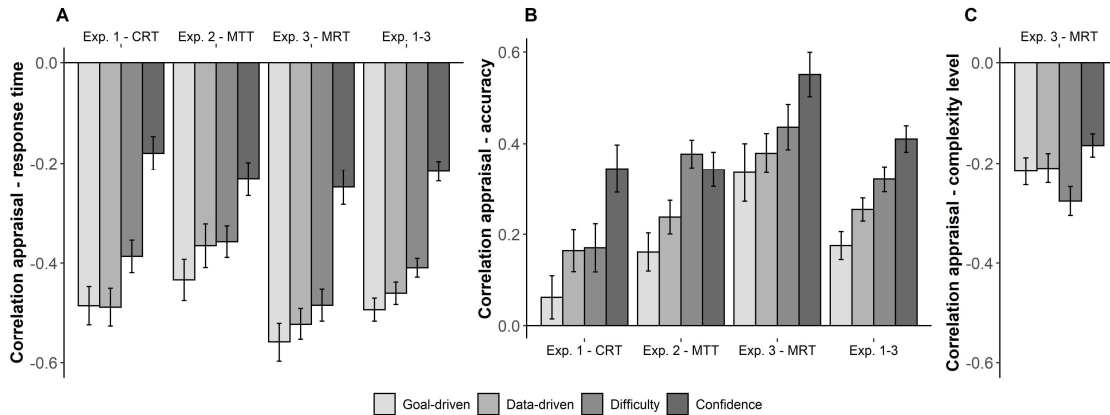
**Table 2**

*Means and Standard Deviations of the Appraisal–Response Time, Appraisal–Accuracy, and Appraisal–Item Complexity Associations as a Function of Type of Appraisal in Experiments 1–3 and Across All Three Experiments*

	Goal-driven effort	Data-driven effort	Task difficulty	Confidence
<u>Experiment 1 (CRT)</u>				
Appraisal–Response time	.49 (.32) ***	.49 (.32) ***	.39 (.27) ***	-.18 (.28) ***
Appraisal–Accuracy	-.06 (.39)	-.16 (.39) **	-.17 (.43) **	.34 (.44) ***
<u>Experiment 2 (MTT)</u>				
Appraisal–Response time	.43 (.30) ***	.37 (.34) ***	.36 (.24) ***	-.23 (.24) ***
Appraisal–Accuracy	-.16 (.31) ***	-.24 (.29) ***	-.38 (.23) ***	.34 (.27) ***
<u>Experiment 3 (MRT)</u>				
Appraisal–Response time	.56 (.29) ***	.52 (.25) ***	.48 (.25) ***	-.25 (.26) ***
Appraisal–Accuracy	-.34 (.46) ***	-.38 (.33) ***	-.44 (.38) ***	.55 (.37) ***
Appraisal–Item complexity	.22 (.21) ***	.21 (.23) ***	.28 (.23) ***	-.16 (.17) ***
<u>Averaged across Experiments 1–3</u>				
Appraisal–Response time	.49 (.31) ***	.46 (.31) ***	.41 (.26) ***	-.21 (.27) ***
Appraisal–Accuracy	-.18 (.40) **	-.25 (.35) ***	-.32 (.38) ***	.41 (.38) ***

*Note.* Associations are calculated as within-participant correlations, and thus range from -1 to 1. Goodman and Kruskal's  $\gamma$  was used to determine the relationship between appraisal and accuracy. Pearson correlations were used for the relationship between appraisal and response time, as well as appraisal and item complexity. Bonferroni adjustment was applied to  $p$ -values. Asterisks indicate whether the averaged correlations significantly differ from 0:  $^+ p < .1$ ,  $^* p < .05$ ,  $^{**} p < .01$ ,  $^{***} p < .001$ .

In accordance with the metacognitive literature, confidence was negatively associated with response time. However, the metacognitive literature suggests that response time can be an unreliable cue for metacognitive judgments (e.g., Ackerman, in press.; Finn & Tauber, 2015). Indeed, as we expected, load-related appraisals seem to rely more strongly on response time as a cue compared to confidence appraisals. Our findings support the idea that load-related appraisals might be biased by response time, as other metacognitive judgments are (see Scheiter et al., 2020).



*Figure 1.* Mean size of correlations of (A) appraisals with response time, (B) appraisals with accuracy, and (C) appraisals with item complexity, as a function of the type of appraisal for the three different tasks in the three experiments (CRT = Cognitive Reflection Task, MTT = Missing Tan Task, MRT = Mental Rotation Task). Note that reversed confidence appraisals were used to calculate the correlations in the confidence group. Error bars show  $\pm 1$  standard error.

***RQ 2: Are There Differences in the Extent to Which Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals Predict Actual Accuracy in Each Task Item?***

To evaluate the predictive value of appraisals for accuracy (see the CRT columns in Figure 1B), the strength of the appraisal–accuracy correlation was compared between the different appraisal types using ANOVA (confidence appraisals were reversed for this comparison). There was a significant medium effect of appraisal type,  $F(3, 272) = 5.61$ ,  $MSE = 0.17$ ,  $p < .001$ ,  $\eta^2 = .06$ . Post-hoc pairwise comparisons ( $t$ -tests with Bonferroni adjustment for  $p$ -values) showed that the correlation with accuracy was significantly stronger for confidence than for goal-driven effort appraisals,  $p < .001$ . All other comparisons showed no significant differences, all  $p_s > .05$ . Thus, while the findings for response time showed a clear distinction between confidence and the load-based appraisal types, at least in the CRT, this distinction was weaker when considering the association with accuracy. Confidence differed from goal-driven effort, but not from the other appraisals.

These predictive values of appraisals for accuracy align with the argument that goal-driven appraisals can reflect labor-in-vain—effort that does not yield improvement (Koriat, 2006), resulting in a weaker relationship with accuracy, compared to confidence appraisals. In this case, as difficulty appraisals shared a similar relationship with accuracy relative to data-driven effort and confidence, it appears to have been interpreted by participants to

reflect the difficulty of the task, rather than the effort they chose to invest in a goal-driven manner.

## Experiment 2

Experiment 2 was designed to replicate the findings of Experiment 1 with a different task. Here, a non-verbal problem-solving task was applied to examine the robustness of the results across different tasks.

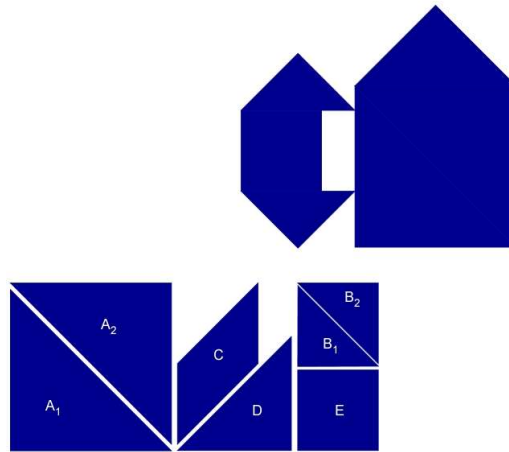
### Method

#### *Participants and Design*

As in Experiment 1, data were collected via Prolific with the same requirements and compensation for participation. Data were excluded for participants who encountered technical problems during the experiment (2 participants); who did not follow the instructions or showed signs of low effort (e.g., failure in four very easy verification items; 15); who provided data for less than 75% of the trials (6); or who did not consent to the use of their data (1). Our final sample comprised 224 participants (age:  $M = 31.9$ ,  $SD = 10.5$ ; 127 female, 94 male; age and gender missing for 3 participants). Again, as in Experiment 1, participants were randomly assigned to one of four groups: goal-driven effort ( $n = 54$ ), data-driven effort ( $n = 60$ ), task difficulty ( $n = 60$ ), and confidence ( $n = 50$ ).

#### *Materials: Missing Tan Task (MTT)*

In the MTT (Ackerman, in press) the silhouette of a figure was shown together with a legend including seven geometric pieces: a square, a parallelogram, two large triangles, two small triangles, and one intermediate triangle (see Figure 2 for an example). The seven geometric pieces were marked with letters from A to E, with two each of the A and B pieces (the large and small triangles respectively). The silhouette was generated from six of the seven pieces, which could be rotated in either direction or flipped to their mirror image but could not overlap. The task was to identify which of the seven geometric pieces was not needed to reproduce the silhouette through mental visualization alone, without being able to physically manipulate or move the pieces. Responses were chosen from five multiple-choice options corresponding to the labels (A to E).



*Figure 2.* An example of the Missing Tan Task from Ackerman (in press). Participants were asked to indicate which of the pieces (A to E) does not fit in the silhouette. In this example, the correct answer is C since all other pieces are needed to reproduce the given silhouette.

We used the original stimuli generated, piloted, and selected by Ackerman (in press), with a total of 32 silhouettes. Two of the items were used as examples during the task instructions. Two other items expected to be easier than the others, with success rates > 90%, served for attention verification, but were included in the analyses, nevertheless. Thus, 30 items were used for the analyses for each participant.

### ***Appraisals***

The appraisals (goal-driven effort, data-driven effort, task difficulty, and confidence) and appraisal procedures were the same as those for Experiment 1. The only exception was that confidence was provided on a different scale (20 to 100, rather than 0 to 100). This change was intended to draw participants' attention to the fact that with five multiple-choice options, they had a 20% chance of being correct just by guessing.

### ***Objective Measures***

As in Experiment 1, response time and accuracy were examined as objective measures that may relate to the appraisals.

### ***Background Variables***

Participants were asked to provide a single-item overall judgment of their performance in the MTT ("How many problems do you think you answered correctly?", from 0 to 30). A single item was also used to elicit participants' experience with solving puzzles ("How often do you solve puzzles or play thought-provoking games?") on a scale from 1 (*almost never*) to 4 (*every day*).

Test anxiety was assessed with five statements about how the participant generally feels about exams (e.g., “During tests I feel very tense.”; Taylor & Deane, 2002). Participants were asked to rate these statements on a 4-point scale (1 = *almost never*, 4 = *almost always*; Cronbach’s  $\alpha = .87$ ). High values on this scale indicate more substantial test anxiety. Participants were also asked when they last took an exam, using a 4-point scale (1 = *during the last month*, 2 = *between 1 and 6 months ago*, 3 = *between 6 and 12 months ago*, and 4 = *more than 12 months ago*).

To assess participants’ mindsets about the malleability of intelligence (fixed vs. growth), they were asked to rate their agreement with four statements (e.g., “You can always substantially change how intelligent you are”; Dweck et al., 1995) on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*; Cronbach’s  $\alpha = .88$ ), with low values indicating a fixed mindset and high values indicating a growth mindset.

One item serving as an attention check was presented together with the test anxiety questions, using the same scale (“Please describe how you generally feel regarding exams: Physical activity promotes my thinking skills. Please ignore this statement and answer by level two”). Less than 6% of participants failed to pass the attention check. However, as in Experiment 1, these participants were only excluded if they showed another indication of inattention (failure to follow instructions or signs of low effort; see under Participants and Design).

Means and standard deviations of the background variables as a function of the type of appraisal are shown in Appendix B.

### ***Procedure***

The overall procedure was similar to that of Experiment 1. In the specific instructions, participants were told their task was to identify which of seven geometric pieces was not needed to reproduce the silhouette by clicking on the corresponding letter. As in Experiment 1, instructions were given for the task with one example, and a second example was used to introduce the appraisals. Participants then worked in a self-paced manner on the 30 MTT items, which were presented randomly. After every ten items, they were told the number of items already completed. Background questions were presented at the end. The full experiment took about 25 minutes.

### ***Data Preparation***

As in Experiment 1, we excluded trials with extraordinarily short (RT < 2 sec; 89 trials) or long response times (RT > 180 sec; 68 trials), as well as trials in which participants



left the experimental environment to do something else (61 trials). This left 6,574 trials to be analyzed. Data preparation and all analyses were performed as in Experiment 1.

### **Results and Discussion**

Item success rates ranged from 8.3% to 81.2%. On average, participants needed 26.1 sec ( $SD = 15.0$ ) to answer each item and were able to solve 43.1% ( $SD = 16.0$ ) of the items correctly. Means and standard deviations of appraisals, response time, and accuracy as a function of the type of appraisal are shown in Table 1. The four appraisal groups did not differ in either response times or accuracy, both  $F < 1$ .

As in Experiment 1, the relationships between the appraisals and the response time and accuracy variables are given as mean within-participant correlations. As seen in Table 2, all mean correlations differed significantly from 0, indicating that overall, response time served as a cue for the appraisals, and appraisals predicted accuracy in the task.

#### ***RQ 1: Are There Differences in the Extent to Which Response Time Serves as a Cue for Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals?***

To evaluate whether the different appraisals rely on response time in similar ways (see the MTT columns in Figure 1A), the strength of the appraisal–response time correlation was compared between the different appraisal types using ANOVA. There was a significant medium effect of appraisal type,  $F(3, 220) = 4.51$ ,  $MSE = 0.08$ ,  $p < .005$ ,  $\eta^2 = .06$ . Post-hoc pairwise comparisons showed that the correlation with response time was significantly weaker for confidence than for goal-driven effort appraisals,  $p = .002$ . All other comparisons showed no significant differences, all  $p_s > .05$ . Thus, results from Experiment 1 were partly replicated, as response time served as a cue for goal-driven effort more than for confidence.

These findings serve an important contribution to the emergent research on non-verbal tasks in the metacognitive literature. In accordance with the scarce studies that investigated non-verbal tasks (e.g., Lauterman & Ackerman, 2019; Reber et al., 2008), the relationship of all appraisals with response time, as well as the replication of the differences between goal-driven and confidence appraisals, demonstrate both shared and distinctive metacognitive mechanisms between verbal and non-verbal tasks.

#### ***RQ 2: Are There Differences in the Extent to Which Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals Predict Actual Accuracy in Each Task Item?***

To evaluate the predictive value of appraisals for accuracy (see the MTT column in Figure 1B), the strength of the appraisal–accuracy correlation was compared between the different appraisal types using ANOVA. There was a significant medium to large effect of

appraisal type,  $F(3, 220) = 7.19$ ,  $MSE = 0.08$ ,  $p < .001$ ,  $\eta^2 = .09$ . Post-hoc pairwise comparisons showed that the correlation with accuracy was significantly stronger for confidence than for goal-driven effort appraisals,  $p = .005$ . In addition, the correlation was significantly stronger for task difficulty appraisals compared with both goal-driven effort,  $p < .001$ , and data-driven effort appraisals,  $p = .036$ . All other comparisons showed no significant differences, all  $p_s > .05$ . In the MTT, similarly to the CRT, confidence was more strongly related to accuracy than goal-driven effort appraisals. However, unlike in Experiment 1, in the MTT, task difficulty appraisals were more strongly related to accuracy than both types of effort appraisals.

Together with the findings regarding response time, we argue that the different appraisals we considered may reflect a continuum in term of how individuals interpret what is asked of them to report (i.e., reflection of self-regulated effort vs. reflection of task demands), in which goal-driven appraisals and confidence appraisals are the two extremes.

### Experiment 3

Experiment 3 was designed to replicate Experiment 1 and Experiment 2 using yet another task, the Mental Rotation Task (MRT). This task offers inherent variations in item complexity by the variation of the angle of rotation between the original shape and its rotated copy, presented among the answer options (see Figure 3 and Materials section below). This task feature allowed us to examine RQ3 as well as RQ1 and RQ2.

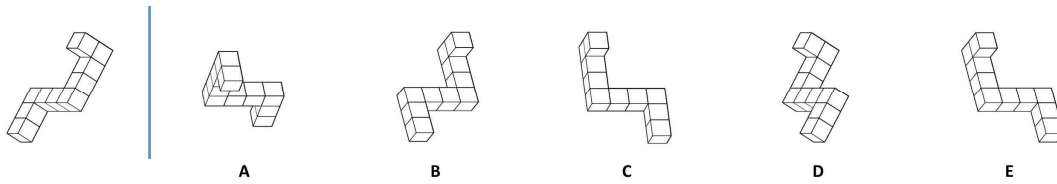
Moreover, recent metacognitive studies have found that people incorporate several cues into their metacognitive judgments at once in both memorizing and problem-solving contexts (Ackerman, in press; Undorf & Bröder, 2020). In particular, these studies encouraged researchers to identify specific task characteristics that predict either success and/or metacognitive judgments for explaining sources for difficulty. Thus, in this experiment we considered the extent to which complexity is taken into account in each of the four appraisals.

### Method

#### *Participants and Design*

Again, we collected data from respondents using Prolific. Participation was remunerated with 3.25GBP. Data were excluded from participants who did not follow instructions (3 participants), had little variability in appraisals (6), completed less than 75% of the trials (4), or did not consent to the use of their data (1). This left 238 participants for the analyses (age:  $M = 26.5$ ,  $SD = 8.3$ ; 81 females, 156 males; age and gender missing for 1

participant). As in the other experiments, participants were randomly assigned to one of four groups that differed in the appraisal they were asked to give: goal-driven effort ( $n = 57$ ), data-driven effort ( $n = 60$ ), task difficulty ( $n = 63$ ), or confidence ( $n = 58$ ).



*Figure 3.* An example of the Mental Rotation Task (items were generated with the Mental Rotation Stimulus Library: Peters & Battista, 2008). Participants were asked to indicate which of the five alternatives (A to E) was identical to the criterion figure on the left. In this example, answer B is correct because it is the same figure but rotated by  $80^\circ$ .

***Materials: Mental Rotation Task (MRT)***

The MRT version used in this study was based on the mental rotation test from Vandenberg and Kuse (1978). We used the mental rotation figures of the Shepard and Metzler type (taken from the Mental Rotation Stimulus Library: Peters & Battista, 2008), comprising three-dimensional line drawings of cubes that are put together to form a figure. Each item consisted of one criterion figure to the left and five alternatives to the right (see Figure 3 for an example). One of the presented answer options was identical to the criterion figure in structure but was shown in a rotated position around either the horizontal or vertical axis. The other alternatives (distractors) were all mirrored versions of the criterion figure which were also rotated around one of the two axes. In addition, the task allowed for systematic manipulation of task complexity at the item level, operationalized as the angle of rotation (angular disparity) between the criterion and the target figure. Previous research showed a linear relation between rotation angle and response time (Shepard & Metzler, 1971). Accordingly, a smaller rotation angle was considered less complex because it entailed less mental rotation, and therefore less time was required to solve the item. We employed nine levels of complexity in steps of  $20^\circ$ , ranging from level 1 with  $20^\circ$  rotation (low complexity) to level 9 with  $180^\circ$  rotation (high complexity). In total, 72 items were used (8 items for each level of complexity), divided into two 36-item sets with 4 items per level in each. Each participant was randomly allocated one of the two sets.

### ***Appraisals***

The goal-driven effort, data-driven effort, task difficulty, and confidence appraisals were elicited as in the other experiments. As in Experiment 2, the confidence appraisal was assessed on a scale from 20 to 100.

### ***Objective Measures***

Again, the objective measures examined were response time and accuracy.

### ***Background Variables***

Overall judgment of performance, experience with puzzle tasks, test anxiety, time since the last exam, and mindset about the malleability of intelligence (fixed vs. growth mindset), were assessed as control variables using the same questions as in Experiment 2. Their means and standard deviations are shown in Appendix B. In addition, an attention check was performed using the same procedure as in Experiment 2. As before, those who did not pass the attention check (less than 6% of the sample) were only excluded if they showed another indication of inattention (see under Participants and Design).

### ***Procedure***

The overall procedure was the same as in the other experiments. In the specific instructions, participants were told their task was to decide which of the five objects shown to the right was the same (rotated) object as the one to the left. The entire task took about 25 minutes.

### ***Data Preparation***

Trials with extraordinarily short ( $RT < 2$  sec; 82 trials) or long response times ( $RT > 180$  sec; 48 trials) were excluded, as were trials in which no response was recorded because of technical problems (42 trials). The final sample comprised 8,448 trials. Data preparation and all analyses were done as in the previous experiments. Complexity (at the item level) is measured on an interval scale, as complexity level corresponds to angular disparity. Hence, Pearson's correlations were used to calculate the correlation between item complexity and appraisals.

### **Results**

Item success rates ranged from 45.6% to 89.7%. On average, participants needed 21.9 sec ( $SD = 11.5$ ) per item, and correctly solved 73.1% ( $SD = 22.5$ ) of the items. Means and standard deviations of appraisals, response time, and accuracy as a function of the type of appraisal are shown in Table 1. The four appraisal groups did not differ in their response times,  $F < 1$ , or in accuracy,  $F(3, 234) = 1.09$ ,  $MSE = 0.05$ ,  $p = .354$ ,  $\eta^2 = .01$ . As seen in

Table 2, all mean correlations significantly differed from 0, indicating that both response time and item complexity served as cues for the appraisals, while the appraisals predicted accuracy in the task.

***RQ 1: Are There Differences in the Extent to Which Response Time Serves as a Cue for Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals?***

To evaluate whether the different appraisals rely similarly on response time (see the MRT columns in Figure 1A), the strength of the appraisal–response time correlation was compared between the different appraisal types using ANOVA. There was a significant large effect of appraisal type,  $F(3, 234) = 16.41$ ,  $MSE = 0.07$ ,  $p < .001$ ,  $\eta^2 = .17$ . Post-hoc pairwise comparisons showed that the correlation with response time was significantly weaker for confidence than for all three of the other appraisal types, all  $p < .001$ , while all other comparisons showed no significant differences, all  $p_s > .05$ . These findings replicate those from Experiment 1 and partly replicate those from Experiment 2, providing further evidence that response time serves as a cue for mental effort and task difficulty appraisals more strongly than confidence appraisals.

***RQ 2: Are There Differences in the Extent to Which Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals Predict Actual Accuracy in Each Task Item?***

To evaluate the predictive value of appraisals for accuracy (see the MRT columns in Figure 1B), the strength of the appraisal–accuracy correlation was compared between the different appraisal types using ANOVA. There was a significant small to medium effect of appraisal type,  $F(3, 224) = 3.25$ ,  $MSE = 0.15$ ,  $p = .023$ ,  $\eta^2 = .04$ . Post-hoc pairwise comparisons showed that the correlation with actual accuracy was significantly stronger for confidence than for goal-driven effort appraisals,  $p < .024$ , while all other comparisons showed no significant differences, all  $p_s > .05$ . Thus, the present results are consistent with those of the previous experiments, in that confidence appraisals were again more strongly related to accuracy than goal-driven effort appraisals.

***RQ 3: Are There Differences in the Extent to Which the Complexity of the Problem Serves as a Cue for Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals?***

First, correlations between complexity and accuracy, as well as complexity and response time, were tested to check the operationalization of item-level complexity. There was a significant negative correlation (Goodman-Kruskal's  $\gamma$  correlation) between complexity and accuracy,  $r = -.11$ ,  $Z = -7.03$ ,  $p < .001$ , indicating that less complex items were more likely to be solved. Furthermore, there was a significant positive correlation

(Pearson's correlation) between complexity and response time,  $r = .09$ ,  $t(8446) = 8.08$ ,  $p < .001$ , indicating that it took participants longer to solve more complex items. Thus, we can conclude that the operationalization of item-level complexity was successful.

To evaluate whether appraisals rely similarly on complexity (see Figure 1C), the strength of the appraisal–complexity correlation was compared between the different types of appraisals using ANOVA. There was a significant small to medium effect of appraisal type,  $F(3, 234) = 2.87$ ,  $MSE = 0.04$ ,  $p = .037$ ,  $\eta^2 = .04$ . Post-hoc pairwise comparisons showed that the correlation with complexity was significantly weaker for confidence than for difficulty appraisals,  $p < .001$ , while all other comparisons showed no significant differences, all  $p_s > .05$ .

In sum, Experiment 3 replicates in another non-verbal problem-solving task our findings regarding response time being a cue for load-related appraisals, as it is for confidence appraisals. It also provides further evidence to the notion that goal-driven effort and confidence lie at opposite ends of a continuum. However, this experiment also has a unique contribution, as it uncovers an additional cue that may guide load-related appraisals in the form of complexity. Although research has shown that all load-related appraisals and confidence are sensitive to task demands variations, our findings suggest that difficulty appraisals rely more strongly on complexity than confidence appraisals do. Notably, this finding may also shed more light on how individuals interpret the difficulty item appraisals: both difficulty and confidence appraisals seem to be perceived as relating to task demands more than individual effort one chooses to invest in the task. However, it may be that difficulty appraisals are a purer reflection of task demands associated with differences in item complexity compared to confidence appraisals.

### **Global Effects: Integrating Results from the Three Experiments**

The results were remarkably consistent across the three experiments, with only slight variations in the findings. However, there were several differences between the tasks. First, while Experiment 1 utilized a verbal task, Experiment 2 and Experiment 3 relied on non-verbal figural tasks. Second, the tasks varied in difficulty, with the mean success rate ranging from 40% in Experiment 2 to 73% in Experiment 3. There were also slight framing variations between the tasks. In particular, while the CRT task had four answer options, the MTT and MRT tasks had five; and the scale used for confidence appraisals ranged from 0 to 100 in the CRT, but 20 to 100 in the MTT and MRT. Thus, apart from the analyses we conducted for each experiment, we also conducted aggregated analyses (again using

ANOVA) to validate our findings beyond these differences and illuminate global effects across the tasks. As seen in Table 2, all correlations significantly differed from zero, indicating that response time served as a cue for all appraisals, and that all appraisals predicted actual accuracy in the task.

***RQ 1: Are There Differences in the Extent to Which Response Time Serves as a Cue for Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals?***

To evaluate whether the different types of appraisals similarly rely on response times (see Exp. 1–3 in Figure 1A), the strength of the appraisal–response time correlation was compared between the different appraisal types across the experiments while controlling for the experimental task. A two-way ANOVA was calculated with appraisal type and experimental task (CRT, MTT, MRT) as between-subject factors, and the appraisal–response time correlation as the dependent variable. There was a significant large main effect of appraisal type,  $F(3, 734) = 34.78$ ,  $MSE = 0.08$ ,  $p < .001$ ,  $\eta_p^2 = .13$ , and a significant small effect of experimental task,  $F(2, 734) = 8.81$ ,  $p < .001$ ,  $MSE = 0.08$ ,  $\eta_p^2 = .02$ . Most importantly, the interaction was not significant,  $F(6, 734) = 1.17$ ,  $p = .318$ ,  $MSE = 0.08$ ,  $\eta_p^2 = .01$ , indicating that the pattern of results was similar across the tasks. Tukey multiple comparisons for appraisal type showed a significantly weaker correlation with response time for confidence than for all the other appraisals, all  $p_s < .001$ , and a weaker correlation for task difficulty appraisals than for goal-driven effort appraisals,  $p = .023$ ; the other comparisons showed no significant differences, all  $p_s > .05$ . Thus, the global pattern suggests that response time is a more reliable cue for effort and difficulty appraisals than for confidence appraisals.

***RQ 2: Are There Differences in the Extent to Which Mental Effort Appraisals, Difficulty Appraisals, and Confidence Appraisals Predict Actual Accuracy in Each Task Item?***

To evaluate the predictive value of appraisals for accuracy (see Exp. 1–3 in Figure 1B), the strength of the appraisal–accuracy correlation was compared between the appraisal types while controlling for the experimental task. Again, a two-way ANOVA was conducted. There were significant medium main effects of appraisal type,  $F(3, 716) = 13.07$ ,  $MSE = 0.13$ ,  $p < .001$ ,  $\eta_p^2 = .05$ , and experimental task,  $F(2, 716) = 26.54$ ,  $MSE = 0.13$ ,  $p < .001$ ,  $\eta_p^2 = .07$ . As in RQ1, the interaction was not significant,  $F < 1$ , indicating that the pattern of results was similar across the tasks. Tukey multiple comparisons for appraisal type showed accuracy to be correlated significantly more strongly with confidence compared with both types of effort appraisals (goal-driven and data-driven), both  $p < .001$ . Accuracy was

also correlated significantly less with goal-driven effort appraisals compared with task difficulty appraisals,  $p < .001$ , while the other comparisons showed no significant differences, all  $ps > .10$ . Thus, overall, confidence appraisals predicted accuracy in the task more reliably than effort appraisals.

### General Discussion

Educational research often relies on people's self-reported subjective appraisals of their experience with learning or problem-solving tasks. However, evidence from both the metacognitive and CLT research domains indicates that such subjective appraisals are prone to biases. Those findings highlight the need for a systematic investigation of the inference processes underlying different appraisals (Scheiter et al., 2020). In the present study, we integrate these two major theoretical approaches, examining the underlying bases and the predictive value of mental effort, task difficulty, and metacognitive confidence appraisals in three cognitively demanding problem-solving tasks by using metacognitive concepts, paradigms, and measures.

Our first research question concerned response time as a cue for subjective appraisals. We were particularly interested in whether response time would emerge as a cue for the load-related appraisals, namely effort and difficulty, as it has for confidence appraisals (e.g., Baars et al., 2020). Across all experiments, we found that, indeed, response time was significantly associated with all appraisals. Yet, interestingly, these associations were stronger for the load-related appraisals than for confidence: the former had moderate to strong relationships with response time, and the latter only a weak relationship.

Using response time as a cue can be informative for mental effort: research has shown that invested effort is related to time investment (Baars et al., 2020). However, metacognitive research has robustly shown that using response time inflexibly as a cue for confidence appraisals in problem-solving tasks can be misleading (e.g., Finn & Tauber, 2015; Thompson, Evans, et al., 2013; Thompson & Morsanyi, 2012). For example, Ackerman and Zalmanov (2012) examined the association between response time and confidence in problem-solving tasks in conditions where response time was either a valid or invalid predictor of performance. Across conditions and regardless of its validity, confidence varied as a function of response time, with participants reporting more confidence in solutions provided quickly than in those which took longer. As mentioned above, recent studies encourage considering multiple sources for actual and perceived difficulty, beyond response time (Undorf & Bröder, 2020). Notably, even when controlling for various such



sources, still, response time as a cue has been shown to result in a monitoring bias (Ackerman, in press). Under the metacognitive framework, such biased monitoring of learning is problematic, as it may misguide subsequent regulatory decisions (e.g., Metcalfe & Finn, 2008b). It remains to be seen whether such harmful effects on subsequent regulatory decisions also arise from bias in load-related appraisals. This question opens fertile ground for follow-up research.

In addition to testing the power of response time to influence subjective appraisals, we also examined the effect of item-level complexity (our third research question). We found support for item complexity as a cue for perceived difficulty, which had a stronger correlation with complexity compared with confidence appraisals but was similar with the two effort appraisals. This finding is in line with the assumption that our operationalization of complexity, as incrementally larger (or smaller) rotation angles, did indeed impose different levels of additional cognitive load, by requiring incrementally more (or less) mental rotation to correctly solve the item. Future research might extend this investigation to other sources of load and to other tasks for delving further into the question of why complexity was found here to be a significantly stronger cue only for difficulty compared with confidence, and not for either of the effort appraisals.

Notably, response time and item complexity are only two of the many cues already uncovered as underlying metacognitive appraisals. Other cues for confidence appraisals in problem-solving tasks identified in the meta-reasoning literature (see Ackerman, 2019 for a review) include accessibility (the number of associations that come to mind when answering a question, e.g., Ackerman & Beller, 2017), self-consistency (the consistency with which different considerations support the chosen answer Bajšanski et al., 2019), and cardinality (the number of considered answer options, Bajšanski et al., 2019). In addition, recent research indicates that cue integration, namely exposing and analyzing multiple cues inherent in the task, has the potential to afford a more thorough understanding of the mechanisms underlying metacognitive appraisals (e.g. Ackerman, in press; Undorf et al., 2018). We call future research to use our methodology to examine other cues and their potential interactive role in load-related appraisals.

Our second research question focused on the predictive value of the various appraisals for item-level success (correct answers). While all appraisals were significantly associated with success, the strength of this association was stronger for both confidence and difficulty appraisals (which were similar, with moderate to large effects) than for effort

appraisals (small to medium effects). Taken together with our finding (RQ1) that response time was more strongly correlated with the effort appraisals than with difficulty or confidence, these findings support the notion that subjective mental effort appraisals, whether goal-driven or data-driven, reflect fluency, and are therefore not a good basis for predicting actual success compared to difficulty and confidence appraisals in the examined tasks. Moreover, Ackerman (in press) succeeded in improving success and attenuating biases in confidence judgments with instructions. It is worth investigating whether such instructional design features affect cues underlying effort appraisals as well.

Overall, our findings support questioning the reliability of the commonly used load-related appraisals as reflections of cognitive load. In addition, the results indicate an important distinction between effort and difficulty appraisals. It has been suggested in CLT research that these scales can be used interchangeably, under the assumption that despite their varied phrasing, they all reflect cognitive load differences stemming from instructional procedures (e.g., de Jong, 2010; Sweller et al., 2011). Our findings suggest that effort appraisals are less accurate than difficulty appraisals in predicting performance. Schmeck et al. (2015) similarly distinguished between effort and difficulty appraisals in relation to performance. In two experiments, they compared single delayed mental effort and difficulty appraisals at the end of a series of tasks to the average of mental effort and difficulty appraisals after each of those tasks. They found that performance was predicted only by mental effort appraisals in one experiment, while difficulty appraisals were more strongly (though not significantly) associated with performance in the other experiment. Our findings, along with those of Schmeck, offer empirical support for the notion that the two measurements reflect distinct constructs (Ayres & Youssef, 2008; van Gog & Paas, 2008).

This study also contributes to the developing field of meta-reasoning (Ackerman & Thompson, 2017). Specifically, very little is known about cue utilization and the predictive value of appraisals in non-verbal reasoning tasks (see Lauterman & Ackerman, 2019). Here we show initial evidence for similar patterns of relationships linking confidence with response time and accuracy in both a well-studied verbal task (the CRT) and two non-verbal tasks (the MTT and MRT). Future studies are called to shed more light on these relationships by examining different tasks, variations in instructional materials, and different populations, while using preregistered hypotheses.

Notably, the present study focused on reasoning and problem-solving tasks. However, both metacognitive and CLT research have been grounded in more typical

learning contexts (e.g., memorization and comprehension tasks). Although similarities in monitoring appraisals have been demonstrated between these different cognitive processes (Ackerman, 2019), it is imperative to examine the generalizability of our findings to more typical learning situations by replicating the study with such tasks, and in actual educational settings. Boundary conditions, such as prior knowledge, also need to be exposed. Finally, further insights into people's reasoning when making appraisals could be gleaned by assessing process data, for instance through think-aloud studies.

### **Limitations**

A first possible limitation of our research is that our samples consisted solely of crowd workers who performed the experiments online. This could have affected the measured response times. However, there is little evidence in the data that the experiments are problematic in that respect. In addition, Prolific is an online research platform that has been empirically found to provide access to more diverse and naïve and less dishonest populations, producing higher data quality with less noise compared to other research platforms (e.g., Gupta et al., 2021; Peer et al., 2017, 2021). Nonetheless, replications under more controlled conditions in the laboratory may serve to reaffirm our findings.

Second, we selected tasks designed to have a certain range of difficulty, where additional effort would improve performance. That is, we chose our tasks so that they should be solvable given sufficient time. Future research should also consider tasks where additional effort does not necessarily pay off in improved performance. These might be simpler tasks in which additional effort merely increases efficiency but not performance or more difficult tasks that participants might not be able to solve even by expending effort.

Third, we used a 0–100 scale to assess the appraisals. While this is common practice in the metacognitive literature, cognitive load research typically employs 5-, 7-, or the original 9-point scales to assess mental effort (Paas, 1992; Paas et al., 2003). On the one hand, a scale from 0 to 100 offers more sensitivity in detecting changes in participants' appraisals. On the other hand, in cognitive load research, it is a subject of debate whether one can distinguish between even nine levels of effort (Paas et al., 2003).

Further, one might argue that our results are limited because of the correlational nature of our approach. Of course, correlational research has limitations like the third-variable problem or that correlations only describe relationships, but causality cannot be inferred. However, it has to be noted that correlations themselves are not the core of our analyses. Rather, we use within-person correlations as dependent variables in an

experimental design. Our interest lies in the varying strength of the relation of varying appraisals with response time, accuracy, or complexity level. Thus, correlations are used as dependent variables in subsequent analyses. These final analyses are the comparison of the correlations as a function of the experimentally varied type of appraisal, which are at the heart of our contribution.

Finally, as noted in the introduction, recent research has begun to delve deeper into developing and validating unique self-report measures for different types of cognitive load— intrinsic, extraneous, and germane load (e.g., Klepsch et al., 2017; Klepsch & Seufert, 2020; Leppink et al., 2013; Leppink & Pérez-Fuster, 2019). While examining these different types of cognitive load was not within the scope of our study, future research could examine how the different types of load relate to response time, performance, and complexity.

### **Conclusion**

While CLT research has assumed that individuals' subjective appraisals of mental load reflect the cognitive resources allocated to achieve task goals, a metacognitive approach suggests that load-related appraisals—like metacognitive appraisals—are potentially susceptible to bias, and in need of thorough investigation (Scheiter et al., 2020). To this end, the present study employs a metacognitive framework to offer novel empirical evidence for the underlying processes on which load-related appraisals rely. The results highlight the tight relationships of load-related appraisals with response time and item-level complexity, as well as their weaker relationship with actual task performance compared to metacognitive confidence appraisals. These findings, which replicate across several unique tasks, imply that load-related appraisals are indeed susceptible to bias.

These findings have powerful implications for research and practice in education. Effective regulation and resource allocation in everyday tasks, and especially in educational contexts, rely heavily on people's ability to accurately appraise the demands of cognitive tasks; and as has been consistently shown in metacognitive research, biased monitoring can mislead future regulatory decisions. Thus, it is important to understand the bases and validity of cues that learners use for effort appraisals and regulation. In practice, the findings can guide the design of different instructional frameworks. For example, designers of adaptive learning environments which rely on self-reports to select the next task should consider which type of appraisal have the strongest associations with their desired learning outcome.

This study should be perceived as a starting point for exposing the underlying processes at the heart of load-related appraisals and to inspire a new stream of future

research. However, the findings already indicate that vigilance is required when collecting and interpreting subjective self-reports of effort and task difficulty. Finally, by relating subjective appraisals of cognitive load to metacognitive appraisals, the present study contributes to bridging the CLT and metacognitive research paradigms (de Bruin et al., 2020; Scheiter et al., 2020).

### References

- Ackerman, R. (in press). Bird's-eye view of cue integration: Exposing instructional and task design factors which bias problem solvers. *Educational Psychology Review*.
- Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*(3), 1349–1368. <https://doi.org/10.1037/a0035098>
- Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments. *Psihologijske Teme*, *28*(1), 1–20. <https://doi.org/10.31820/pt.28.1.1>
- Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgements during reasoning and memorisation. *Thinking & Reasoning*, *23*(4), 376–408. <https://doi.org/10.1080/13546783.2017.1328373>
- Ackerman, R., & Thompson, V. (2015). Meta-reasoning: What can we learn from meta-memory. In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory* (pp. 164–178). Psychology Press.
- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187–1192. <https://doi.org/10.3758/s13423-012-0305-z>
- Ashburner, M., & Risko, E. F. (2021). Judgements of effort as a function of post-trial versus post-task elicitation. *Quarterly Journal of Experimental Psychology*, *74*(6), 991–1006. <https://doi.org/10.1177/17470218211005759>
- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400. <https://doi.org/10.1016/j.learninstruc.2006.09.001>
- Ayres, P., & Youssef, A. (2008). Investigating the influence of transitory information and motivation during instructional animations. In P. A. Kirschner, F. Prins, V. Jonker, & G. Kanselaar (Eds.), *Proceedings of the 8th International Conference for the Learning Sciences* (pp. 68–75). ICLS.
- 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*, *32*(4), 979–1002. <https://doi.org/10.1007/s10648-020-09569-3>

- Bajšanski, I., Žauhar, V., & Valerjev, P. (2019). Confidence judgments in syllogistic reasoning: The role of consistency and response cardinality. *Thinking & Reasoning*, 25(1), 14–47. <https://doi.org/10.1080/13546783.2018.1464506>
- Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Metacognition and implicit memory* (pp. 309–338). Erlbaum.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55–68. <https://doi.org/10.1037/0096-3445.127.1.55>
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64(1), 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>
- Blissett, S., Sibbald, M., Kok, E., & van Merriënboer, J. (2018). Optimizing self-regulation of performance: Is mental effort a cue? *Advances in Health Sciences Education*, 23(5), 891–898. <https://doi.org/10.1007/s10459-018-9838-x>
- Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 53–61. [https://doi.org/10.1207/S15326985EP3801\\_7](https://doi.org/10.1207/S15326985EP3801_7)
- Brünken, R., Seufert, T., & Paas, F. (2006). Measuring Cognitive Load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive Load Theory* (pp. 181–202). Cambridge University Press. <https://doi.org/10.1017/CBO9780511844744.011>
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131. <https://doi.org/10.1037/0022-3514.42.1.116>
- Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, 36(2), 429–437. <https://doi.org/10.3758/MC.36.2.429>
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293–332. [https://doi.org/10.1207/s1532690xci0804\\_2](https://doi.org/10.1207/s1532690xci0804_2)
- Chen, S., Epps, J., & Paas, F. (2022). Pupillometric and blink measures of diverse task loads: Implications for working memory models. *British Journal of Educational Psychology*, 00, 1–21. <https://doi.org/10.1111/bjep.12577>
- de Bruin, A. B. H., Roelle, J., Carpenter, S. K., & Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review*, 32(4), 903–915. <https://doi.org/10.1007/s10648-020->

09576-4

- de Bruin, A. B. H., & van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1–9.  
<https://doi.org/10.1016/j.learninstruc.2017.06.001>
- de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*(2), 105–134.  
<https://doi.org/10.1007/s11251-009-9110-0>
- De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*(2), 269–273.  
<https://doi.org/10.3758/s13423-013-0384-5>
- Dunn, T. L., Gaspar, C., & Risko, E. F. (2019). Cue awareness in avoiding effortful control. *Neuropsychologia, 123*, 77–91. <https://doi.org/10.1016/j.neuropsychologia.2018.05.011>
- Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: Roles of perceived error-likelihood and time demands. *Psychological Research, 83*(5), 1033–1056. <https://doi.org/10.1007/s00426-017-0943-x>
- Dunn, T. L., & Risko, E. F. (2016). Toward a metacognitive account of cognitive offloading. *Cognitive Science, 40*(5), 1080–1127. <https://doi.org/10.1111/cogs.12273>
- Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry, 6*(4), 267–285.  
[https://doi.org/10.1207/s15327965pli0604\\_1](https://doi.org/10.1207/s15327965pli0604_1)
- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self- and co-regulation. *European Psychologist, 13*(4), 277–287.  
<https://doi.org/10.1027/1016-9040.13.4.277>
- 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 University Publishing. <https://doi.org/10.17885/heiup.470.c6669>
- Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27*(4), 567–586.  
<https://doi.org/10.1007/s10648-015-9313-7>
- Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-



- making competence of older adults. *Psychology and Aging*, 25(2), 271–288.  
<https://doi.org/10.1037/a0019106>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.  
<https://doi.org/https://doi.org/10.1257/089533005775196732>
- Gupta, N., Rigotti, L., & Wilson, A. (2021). The experimenters' dilemma: Inferential preferences over populations. *ArXiv Preprint*. <http://arxiv.org/abs/2107.05064>
- Haji, F. A., Rojas, D., Childs, R., de Ribaupierre, S., & Dubrowski, A. (2015). Measuring cognitive load: Performance, mental effort and simulation task complexity. *Medical Education*, 49(8), 815–827. <https://doi.org/10.1111/medu.12773>
- Hawkins, G. E., & Heathcote, A. (2021). Racing against the clock: Evidence-based versus time-based decisions. *Psychological Review*, 128(2), 222–263.  
<https://doi.org/10.1037/rev0000259>
- Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1191–1206.  
<https://doi.org/10.1037/a0013025>
- Jonsson, A.-C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences*, 34(4), 559–574. [https://doi.org/10.1016/S0191-8869\(02\)00028-4](https://doi.org/10.1016/S0191-8869(02)00028-4)
- Kelley, C. M. ., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32(1), 1–24. <https://doi.org/10.1006/jmla.1993.1001>
- Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language*, 35(2), 157–175.  
<https://doi.org/10.1006/jmla.1996.0009>
- 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(August), 101237.  
<https://doi.org/10.1016/j.cogpsych.2019.101237>
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in*

- Psychology*, 8, 1–18. <https://doi.org/10.3389/fpsyg.2017.01997>
- Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, 48(1), 45–77. <https://doi.org/10.1007/s11251-020-09502-9>
- Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, 30(2), 503–529. <https://doi.org/10.1007/s10648-017-9404-8>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory and Cognition*, 36(2), 416–428. <https://doi.org/10.3758/MC.36.2.416>
- Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135(1), 36–69. <https://doi.org/10.1037/0096-3445.135.1.36>
- Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 117–135). Psychology Press.
- Lauterman, T., & Ackerman, R. (2019). Initial judgment of solvability in non-verbal problems – a predictor of solving processes. *Metacognition and Learning*, 14(3), 365–383. <https://doi.org/10.1007/s11409-019-09194-8>
- Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45(4), 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Leppink, J., & Pérez-Fuster, P. (2019). Mental effort, workload, time on task, and certainty: Beyond linear models. *Educational Psychology Review*, 31(2), 421–438. <https://doi.org/10.1007/s10648-018-09460-2>
- Lins de Holanda Coelho, G., H. P. Hanel, P., & J. Wolf, L. (2020). The very efficient assessment of need for cognition: developing a six-item version. *Assessment*, 27(8),

- 1870–1885. <https://doi.org/10.1177/1073191118793208>
- Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, 7(4), 263–269.  
<https://doi.org/https://doi.org/10.1111/j.1745-3984.1970.tb00727.x>
- Metcalfe, J., & Finn, B. (2008a). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1084–1097. <https://doi.org/10.1037/a0012580>
- Metcalfe, J., & Finn, B. (2008b). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15(1), 174–179.  
<https://doi.org/10.3758/PBR.15.1.174>
- Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? *Psychological Science*, 22(3), 320–324.  
<https://doi.org/10.1177/0956797610397954>
- Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, 19(4–5), 650–670. <https://doi.org/10.1080/09541440701326196>
- Naismith, L. M., Cheung, J. J. H., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, 49(8), 805–814. <https://doi.org/10.1111/medu.12732>
- National Institute for Testing and Evaluation. (n.d.). *The psychometric entrance test - practice tests*. <https://www.nite.org.il/psychometric-entrance-test/preparation/?lang=en>
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, Issue C, pp. 125–173). Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia*, 91, 499–508.  
<https://doi.org/10.1016/j.neuropsychologia.2016.09.010>
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434. <https://doi.org/10.1037/0022-0663.84.4.429>
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load

- measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71. [https://doi.org/10.1207/S15326985EP3801\\_8](https://doi.org/10.1207/S15326985EP3801_8)
- Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G., & Aubteen Darabi, A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational Technology Research and Development*, 53(3), 25–34. <https://doi.org/10.1007/BF02504795>
- Paas, F., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351–371. <https://doi.org/10.1007/BF02213420>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8:422, 1–28. <https://doi.org/10.3389/fpsyg.2017.00422>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3>
- Peters, M., & Battista, C. (2008). Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. *Brain and Cognition*, 66(3), 260–264. <https://doi.org/10.1016/j.bandc.2007.09.003>
- Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 318–329). Guilford Press.
- Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, 29(5), 453–469. <https://doi.org/10.1002/bdm.1883>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- 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>

- Reber, R., Brun, M., & Mitterndorfer, K. (2008). The use of heuristics in intuitive mathematical judgment. *Psychonomic Bulletin & Review*, *15*(6), 1174–1178. <https://doi.org/10.3758/PBR.15.6.1174>
- 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>
- Rop, G., Schöler, A., Verkoeijen, P. P. J. L., Scheiter, K., & Gog, T. (2018). Effects of task experience and layout on learning from text and pictures with or without unnecessary picture descriptions. *Journal of Computer Assisted Learning*, *34*(4), 458–470. <https://doi.org/10.1111/jcal.12287>
- Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review*, *32*(4), 1003–1027. <https://doi.org/10.1007/s10648-020-09555-9>
- 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*(1), 93–114. <https://doi.org/10.1007/s11251-014-9328-3>
- Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, *6*(4), 147–151. <https://doi.org/10.1027/1614-2241/a000016>
- Schwartz, B. L., & Jemstedt, A. (2021). The role of fluency and dysfluency in metacognitive experiences. In D. Moraitou & P. Metallidou (Eds.), *Trends and prospects in metacognition research across the life span* (pp. 25–40). Springer International Publishing. [https://doi.org/10.1007/978-3-030-51673-4\\_2](https://doi.org/10.1007/978-3-030-51673-4_2)
- Searle, J. A., & Hamm, J. P. (2017). Mental rotation: An examination of assumptions. *WIREs Cognitive Science*, *8*(6), 701–703. <https://doi.org/10.1002/wcs.1443>
- Seufert, T. (2020). Building bridges between self-regulation and cognitive load—an invitation for a broad and differentiated attempt. *Educational Psychology Review*, *32*(4), 1151–1162. <https://doi.org/10.1007/s10648-020-09574-6>
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. <https://doi.org/10.1126/science.171.3972.701>
- Shtulman, A., & McCallum, K. (2014). Cognitive reflection predicts science understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2937–2942.

- 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>
- Sirota, M., Dewberry, C., Juanchich, M., Kostovičová, L., & Marshall, A. C. (2018). *Measuring cognitive reflection without maths: Developing and validating the verbal cognitive reflection test*. PsyArXiv. <https://doi.org/https://doi.org/10.31234/osf.io/pfe79>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory* (pp. 71–85). Springer New York. [https://doi.org/10.1007/978-1-4419-8126-4\\_6](https://doi.org/10.1007/978-1-4419-8126-4_6)
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Szulewski, A., Kelton, D., & Howes, D. (2017). Pupillometry as a tool to study expertise in medicine. *Frontline Learning Research, 5*(3), 55–65. <https://doi.org/10.14786/flr.v5i3.256>
- Taylor, J., & Deane, F. P. (2002). Development of a short form of the test anxiety inventory (TAI). *The Journal of General Psychology, 129*(2), 127–136. <https://doi.org/10.1080/00221300209603133>
- Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. <https://doi.org/10.1037/0022-0663.95.1.66>
- Thompson, V. A., Evans, J. S. B. T., & Campbell, J. I. D. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning, 19*(3–4), 431–452. <https://doi.org/10.1080/13546783.2013.820220>
- Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society, 11*(1), 93–105. <https://doi.org/10.1007/s11299-012-0100-6>
- Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*(2), 237–251. <https://doi.org/10.1016/j.cognition.2012.09.012>
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information

- processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, 20(2), 147–168. <https://doi.org/10.1080/13546783.2013.844729>
- Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1448–1457. <https://doi.org/10.1037/xlm0000248>
- Undorf, M. (2020). Fluency illusions in metamemory. In *Memory Quirks* (pp. 150–174). Routledge. <https://doi.org/10.4324/9780429264498-12>
- Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, 73(4), 629–642. <https://doi.org/10.1177/1747021819882308>
- Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, 46(4), 507–519. <https://doi.org/10.3758/s13421-017-0780-6>
- Valerjev, P. (2019). Chronometry and meta-reasoning in a modified cognitive reflection test. In K. Damnjanović, O. Tošković, & S. Marković (Eds.), *Proceedings of the XXV Scientific Conference: Empirical Studies in Psychology* (pp. 31–34).
- van Gog, T. (2022). The signaling (or cueing) principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 221–230). Cambridge University Press. <https://doi.org/10.1017/9781108894333.022>
- van Gog, T., Hoogerheide, V., & van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review*, 32(4), 1055–1072. <https://doi.org/10.1007/s10648-020-09544-y>
- van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, 26(6), 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(1), 16–26. <https://doi.org/10.1080/00461520701756248>
- van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, 27(2), 247–264. <https://doi.org/10.1007/s10648-015-9310-x>

- van Merriënboer, J. J. G., & Kirschner, P. A. (2017). *Ten steps to complex learning: A systematic approach to four-component instructional design* (3rd ed.). Routledge.  
<https://doi.org/10.4324/9781315113210>
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599–604.  
<https://doi.org/10.2466/pms.1978.47.2.599>
- Wang, S., & Thompson, V. (2019). Fluency and feeling of rightness: The effect of anchoring and models. *Psihologijske Teme*, *28*(1), 37–72. <https://doi.org/10.31820/pt.28.1.3>
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Academic Press.
- Young, A. G., Powers, A., Pilgrim, L., & Shtulman, A. (2018). Developing a cognitive reflection test for school-age children. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1232–1237). Cognitive Science Society.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. [https://doi.org/10.1207/s15430421tip4102\\_2](https://doi.org/10.1207/s15430421tip4102_2)



## Appendix

**Appendix A – Development and Pretesting of the CRT Items**

Initially, 31 open-ended problems were compiled from different studies and publications (De Neys et al., 2013; Finucane & Gullion, 2010; National Institute for Testing and Evaluation, n.d.; Oldrati et al., 2016; Primi et al., 2016; Shtulman & McCallum, 2014; Sirota et al., 2018; Toplak et al., 2014; Trippas et al., 2016; Valerjev, 2019; Young et al., 2018). Thirty-two participants completed a pretest of these items via Prolific for 2GBP monetary compensation. Items were presented to participants in random order. Following analysis of success rates and response times, we excluded eight items for the following reasons: very high success rates with quick response times; previously known to many participants; very low success rates with long response times; and minimal number of expected answers with long response times. This analysis left us with 23 items. We then fitted the problems to a multiple-choice format. The multiple-choice items included four answer options generated such that each item had a correct answer, one misleading answer (i.e., an answer that was incorrect but predictable), and two distractors. The two distractors were developed based on the pretest, and comprised the two (incorrect) answers provided by the highest proportion of pretest participants. If all given answers appeared at the same rates, we selected those that seemed most misleading. We generated new distractors for five items which did not result in enough wrong answers in the pretest. Thirty-four participants then completed a second pretest using the multiple-choice format. One problem was excluded for being too easy, two problems were selected as training items, and three very easy problems were selected as attention check items, leaving 17 problems in the final set.

**Appendix B – Descriptive Statistics of Background Variables**

*Means and Standard Deviations of Background Variables as a Function of Type of Appraisal in Experiments 1–3*

	Goal-driven effort	Data-driven effort	Task difficulty	Confidence
<u>Experiment 1 (CRT)</u>				
Need for cognition	4.57 (1.25)	4.68 (1.13)	4.96 (1.05)	5.03 (0.86)
Experience with puzzles	4.05 (1.58)	3.84 (1.72)	3.93 (1.60)	4.39 (1.58)
<u>Experiment 2 (MTT)</u>				
Judgment of performance	50.12 (22.51)	46.27 (21.95)	45.23 (21.89)	45.44 (19.35)
Experience with puzzles	1.81 (0.72)	1.81 (0.69)	1.77 (0.71)	1.92 (0.70)
Test anxiety	2.67 (0.81)	2.48 (0.70)	2.62 (0.82)	2.36 (0.79)
Time since last exam	3.29 (1.07)	3.05 (1.00)	2.86 (1.23)	3.10 (1.13)
Mindset of intelligence	4.54 (1.21)	4.37 (1.24)	4.60 (1.31)	4.77 (1.21)
<u>Experiment 3 (MRT)</u>				
Judgment of performance	74.61 (20.64)	70.28 (19.25)	71.21 (21.65)	73.13 (22.40)
Experience with puzzles	2.00 (0.80)	1.87 (0.70)	1.87 (0.68)	2.10 (0.77)
Test anxiety	2.25 (0.68)	2.43 (0.72)	2.37 (0.74)	2.31 (0.80)
Time since last exam	2.51 (1.15)	2.45 (1.08)	2.30 (1.29)	2.52 (1.20)
Mindset of intelligence	4.62 (1.25)	4.42 (1.28)	4.88 (1.14)	4.68 (1.26)

*Note.* Need for cognition was assessed on a 7-point Likert scale, with low scores indicating low need for cognition. Experience with puzzles was assessed on a 7-point scale in Experiment 1 and on a 4-point scale in Experiment 2 and Experiment 3, with low scores always indicating little experience. Judgment of performance was assessed as the number of correctly solved tasks and is given as a percentage of the number of tasks in the relevant experiment. Test anxiety and mindset about the malleability of intelligence were assessed on 4-point scales, with low scores indicating low test anxiety or a fixed mindset, respectively. Time since the last exam was assessed on a scale from 1 (*during the last month*) to 4 (*more than 12 months ago*).