April, 2023

Reference:
Ackerman, R. (2023). Bird's-Eye View of Cue Integration:
Exposing instructional and task design factors which bias problem solvers. *Educational Psychology Review*, *35*:*55*. https://doi.org/10.1007/s10648-023-09771-z
This preprint has not undergone peer review or any post-submission improvements or corrections.

Bird's-Eye View of Cue Integration:

Exposing Instructional and Task Design Factors Which Bias Problem Solvers

Rakefet Ackerman

Technion—Israel Institute of Technology

Author Note

Correspondence concerning this article should be addressed to R. Ackerman.

Faculty of Data and Decision Sciences

Technion—Israel Institute of Technology, Haifa 3200003, Israel

ORCID: https://orcid.org/my-orcid?orcid=0000-0001-9583-8014

EMAIL: ackerman@technion.ac.il

The code and data are available in OSF: https://osf.io/jgn54/?view_only=d66bed9658d14932ab8a1bd86b1dec97

Keywords: Metacognition; Meta-Reasoning; Heuristics and Biases; Problem Solving

Declarations: I am an editorial board member of Educational Psychology Review. This work was supported by the Israel Science Foundation. No financial interests are involved. I would like to thank Meira Ben-Gad for editorial assistance and Paul Feigin for statistical consulting. The research methodologies were approved by the Behavioral Sciences Research Ethics Committee of the Technion—Israel Institute of Technology (approval number 2020-015). Informed consent was collected from all participants at the outset of each experiment that was conducted by the author as part of this research program.

Abstract

Solving problems in educational settings, as in daily-life scenarios, involves constantly assessing one's own confidence in each considered solution. Metacognitive research has exposed cues that may bias confidence judgments (e.g., familiarity with question terms). Typically, metacognitive research methodologies require examining misleading cues one-by-one, while recent research has revealed integration of multiple cues stemming from the same stimuli. However, this research leaves open important questions about including the weight balance among cues and their changes across task design (e.g., instructions) and/or population characteristics (e.g., background knowledge). The present study presents the *Bird's-Eye View of Cue Integration* (BEVoCI) methodology. It is based on hierarchical multiple regression models, allowing efficient exposure of multiple biases at once, their relative weights, and their malleability across task designs and populations. Notably, the BEVoCI can be applied both to planned studies and to existing datasets. I demonstrate its application in both ways. In Experiment 1 and Experiment 2, I introduce two nonverbal problem-solving tasks, the Comparison of Perimeters (CoP) and the novel Missing Tan Task (MTT), while Experiment 3 reanalyzes data collected by others, comprising algebra problems solved by children and adults. The experiments demonstrate exposing biases, their malleability across conditions, and the non-straightforward association between performance improvement and overcoming biases; and the results of Experiment 3 provide strong support for the generalizability of the methodology. Pinpointing sources of bias is essential for guiding educational design efforts.

1. Introduction

When answering knowledge questions in a conversation, in the classroom, during homework, or in an exam, a confidence judgment is the self-assessed chance that each considered answer is correct. These metacognitive judgments of confidence, like judgments of learning, are expected to be in line with actual success so long as the heuristic cues they are derived from are reliable in the specific case at hand (Koriat, 1997). It is well-established that people who are better at monitoring their knowledge have better real-life outcomes than those whose monitoring is less accurate (see Kleitman & Moscrop, 2010). It is also known that reliable metacognitive judgments regarding each task item (e.g., a homework question) are important for effective self-regulated learning (see Bjork et al., 2013; de Bruin et al., 2020; Fiedler et al., 2019). Therefore, among the central aims of educational task design are to identify sources of systematic bias stemming from misleading heuristic cues (e.g., Chen et al., 2018; see Scheiter et al., 2020, for a review); to identify differences in biases between learners and conditions (e.g., Händel et al., 2020); and to help learners overcome these biases (see Sweller et al., 2019, for a review). Knowing the factors that generate biases across designs and contexts is thus a prerequisite for educators' efforts to improve learners' sensitivity to pitfalls.

Importantly, existing methodologies in metacognitive research tend to examine hypothesized biasing cues one-by-one—a limitation that impedes our understanding of how these cues interact, and how their balance changes under different conditions. The same is true for studies that attempt to overcome such biases. These studies typically either deal with one biasing factor—e.g., one item characteristic or task design feature—at a time (e.g., Castel, 2008; Koriat & Bjork, 2006; Yan et al., 2016); or they examine global success and monitoring, but not the effects of eliminating specific cue-related biases (e.g., Ariel et al., 2021; Baars et al., 2014; Raaijmakers et al., 2019; Tauber et al., 2018; Thiede et al., 2022). Thus, these studies overlook the possibility that improvement techniques may overcome some biases but not others and may even generate new biases.

The present study presents a new data analysis approach, the *Bird's-Eye View of Cue Integration* (BEVoCI). This methodology is based on hierarchical multiple regression models, and it offers a fourfold advantage over existing methodologies. First, it allows systematic and efficient analysis of several cues at the same time with one group of participants. Second, it allows exposing the integration of these cues while considering the relative weights of each cue. Third, it efficiently enables exposing biases by comparing how cues are utilized for metacognitive judgments relative to their prediction of an objective measure of success. Finally, Ackerman (2019) differentiated between heuristic cues at the item level (level c), task level (level b), and individual level (level a). Particularly, she pointed out the often-overlooked possibility that such contextual factors may affect how cues are utilized. The BEVoCI methodology enables considering the malleability of cues, their integration, and biases across such contextual characteristics.

The present study deals with the still relatively new research domain of meta-reasoning (Ackerman & Thompson, 2017). So far, our understanding of biases in metacognitive judgments comes mostly from *meta-memory* tasks, involving either memorization or knowledge retrieval (Bjork et al., 2013). Memory tasks involve coming up with a specific target item that participants have encountered before. *Meta-reasoning* research generalizes well-established principles from meta-memory (Ackerman & Thompson, 2015), but adds metacognitive elements unique to reasoning and problem-solving. For instance, the solving process might require applying skills acquired in different contexts (e.g., using math-based thinking to address an engineering challenge). Thus, meta-reasoning processes constantly monitor the current state of the problem-solving process and control self-regulated effort investment (see Ackerman & Thompson, 2017, for a review).

I begin by introducing relevant concepts and classic methods for identifying cues that underlie and may bias metacognitive judgments. I then introduce the BEVoCI data analysis approach, the main contribution of the present study. In the empirical part of the paper, I apply the BEVoCI approach to three different tasks, in order to demonstrate the generalizability of the method and the insights it provides, as well as to inspire future research.

1.1. Cue Validity and Cue Utilization: Identifying Biasing Factors

Brunswik’s lens models are conceptual models accompanied by statistical tools that enable assessing the extent to which a set of information sources (hereafter, “cues”) predict actual performance compared to subjective judgments. These models have been widely applied to teachers’ judgments. For instance, they have been used to discover factors that underlie and might distort teachers' subjective evaluations of students' performance (e.g., bin Mohd Noh & bin Mohd Matore, 2019; Kaufmann, 2022).

Applying this approach to metacognitive judgments, Koriat (1997) differentiated between *cue validity*, the objective predictive value of a cue for success, and *cue utilization*—the effect of the same cue on metacognitive judgments. Using well-defined tasks which have decisive correct responses (e.g., exam questions, rather than flexible design tasks) allows exposing discrepancies between the two, which reflect systematic biases. For instance, Ackerman et al. (2013) compared participants' subjective assessments of how well they understood solution explanations. The examined cue was the presence versus absence of non-informative illustrations. The results showed that when explanations included illustrations, participants reported higher metacognitive judgments of understanding but had lower actual success rates than in the absence of illustrations. Such predictable biases are expected to reduce the efficiency of regulatory decisions (e.g., whether to continue studying) in self-regulated learning and thinking because people base these decisions on distorted self-assessments of their knowledge (e.g., Desender et al., 2018; Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008).

Biases in metacognitive judgments have been found with many types of tasks. For instance, semantically related word pairs (kite – sky) are easier to remember than unrelated ones. In general, this fact is reflected in learners' judgments of learning (JOL; e.g., Mueller et al., 2013; see Undorf et al., 2018). However, the effect of relatedness on JOL was found to be stronger when related and unrelated word pairs were intermixed than when presented in separate lists (Koriat et al., 2009). This context effect represents a level b cue in Ackerman's (2019) taxonomy. In other cases, metacognitive judgments have been influenced by factors that in reality either do not affect success, or influence success and judgments differently (e.g., concreteness, Markovits et al., 2015; overt vs. covert retrieval, Tauber et al., 2018).

One type of bias that is difficult to identify is when people get the direction of a cue effect right, but under- or overestimate its size. An example is font-size: larger fonts (e.g., 48 points) tend to increase judgments of learning relative to smaller fonts (e.g., 18 points). This is the case even though success is often unaffected by font size (Rhodes & Castel, 2008). When a larger font size does increase success, it was found to disproportionately inflate judgments of learning (e.g., Halamish, 2018; Undorf et al., 2018). The BEVoCI method allows us to expose such biases and compare them statistically across conditions.

1.2. Cue Integration: Exposing Multiple Biases at Once

Recent research suggests that people integrate multiple cues in their metacognitive judgments of their chance to succeed. That is, their judgments reflect several task and situation characteristics that may affect success. By definition, cue integration is specific to the confluence of a particular person, task, and context. Thus, in terms of research methodology, cue integration is best studied when a single sample faces a single task. Previous studies have documented extensive integration of 4-5 cues for meta-memory judgments, with all cues varying across items within participants (Undorf & Bröder, 2021; Undorf et al., 2018). Notably, in those studies Brunswik's lens models were used to assess the overall validity of judgments based on the combined set of cues. However, the relative weights of the cues which together underlie the metacognitive judgment remained a black box.

The BEVoCI method offered here allows opening this black box, exposing the relative weights of the cues—i.e., determining which one(s) dominate others—when predicting separately success and confidence (or any other metacognitive judgment) in each condition. It is based on hierarchical multiple linear regression models, which enable identifying the unique contribution of each cue while controlling for other cues. It also supports exposing confidence biases by statistically comparing cue weights between predictions of success and confidence; a significant difference indicates a bias. Three types of biases are possible: (i) a cue predicts either success or confidence, but not both; (ii) it predicts both, but with significantly different weights; or (iii) it predicts both but in opposite directions. Finding several types of biases within one condition points to a double dissociation between success and confidence because they together indicate that different factors uniquely affect each one.

The ultimate aim of the BEVoCI is to help educators identify misleading factors in combinations of task items, task design, and populations. The BEVoCI also aims to inform intervention designers about remaining biases after an improvement attempt. Thus, the BEVoCI supports a realistic assessment of the costs and benefits of intervention programs.

1.3. Overview of the Experiments

As in memory tasks, reasoning and problem-solving research employ mostly verbal stimuli (Ackerman & Thompson, 2017). As a result, evidence for cue utilization has been obtained mainly with semantic cues (e.g., word association nets, Ackerman & Beller, 2017; word concreteness, Markovits et al., 2015). The present study uses two visual reasoning tasks new to metacognitive research, which are based on geometric shapes. Experiment 1 employs an extended version of the Comparison of Perimeters (CoP) task inspired by an educational research methodology introduced by Stavy and Babai (2008). Experiment 2 introduces the Missing Tan Task (MTT), which involves identifying the geometric shapes hidden in silhouettes. It is based on the Tangram game (https://en.wikipedia.org/wiki/Tangram), which has also been used for educational research (Bohning & Althouse, 1997; Lee et al., 2009). Both tasks call for spatial reasoning processes such as those involved in design, navigation, engineering, and when learning geometry (e.g., Hart et al., 2017; Roll et al., 2014; Zak et al., 2021). Relative to verbal reasoning tasks, the roles of vocabulary and semantic knowledge are reduced. In addition, the task versions developed for this study do not allow physical manipulation of shapes. These factors put a large burden on mental imagery and visuospatial working memory (e.g., manipulating mental images; see Bates & Farran, 2021; Castro-Alonso et al., 2019; Träff et al., 2019, for reviews). Thus, the present study contributes the two nonverbal tasks alongside its main contribution—the BEVoCI methodology. These tasks are described in detail in the introduction and method sections of each experiment.

The visual tasks used here allow considering cues known based on past research (e.g., response time, serial order), together with novel cues hypothesized to generate biases specific to such stimuli (e.g., perimeter, number of edges). Exposing unknown misleading cues carries theoretical and practical implications expected to guide future basic and applied research. From a research methodology perspective, the two visual tasks have the following advantages: (i) they can be used without concern for language limitations; (ii) they allow for generating items at a large variety of objective and subjective difficulty levels; (iii) dozens of items can be solved in less than half an hour, supporting robust within-participant statistical analyses; (iv) they support variations in task design and instructions; and (v) they provide opportunities for identifying multiple potentially misleading heuristic cues within the same task.

Clearly, though, the BEVoCI methodology is well-suited to any other task that fits the task choice criteria detailed below. To demonstrate this generalizability, Experiment 3 comprises a reanalysis of a dataset collected by others (Vuorre and Metcalfe, 2022; see below), who employed algebra problems used in school curricula. All three tasks were used in this study in multiple-choice test formats, and thus most invested time was devoted to mental manipulation and thinking.

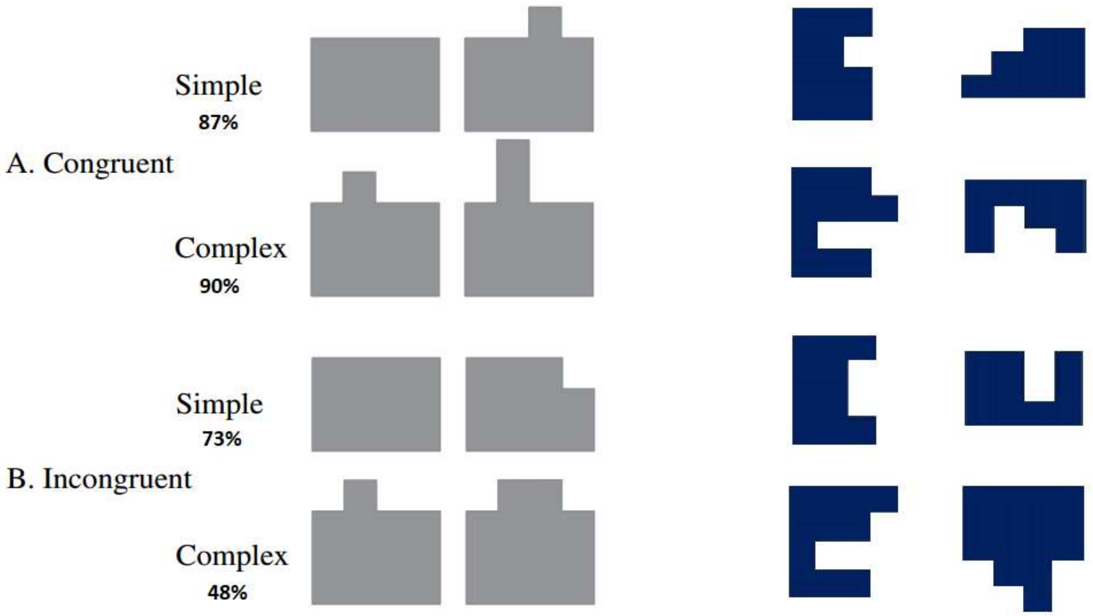
The global dependent variables in all experiments were the classic metacognitive research measures (see Ackerman et al., 2016, for details regarding these measures)—namely success, confidence, response time (thinking time), and several associations between these measures: efficiency (correct responses per minute of work); overconfidence (mean confidence minus participants' success rate as a percentage); and resolution (the within-participant success–confidence gamma correlation, reflecting confidence discrimination between correct and wrong responses). In the algebra dataset, response time was not documented, and thus it and efficiency (correct responses per unit of thinking time) are missing from the analyses. The main dependent variables, though, were the beta coefficients of the various cues examined for each task by the hierarchical multiple regression models, using the BEVoCI method.

2. Experiment 1: Biasing Cues and The Effects of Knowledge and Instruction

This experiment consists of three sub-experiments with five groups altogether, using the Comparison of Perimeters (CoP) task. In the basic version of the task, used in Exp1a, participants were presented with two geometric shapes generated from squares and had to decide which, if either, shape had a longer perimeter. The CoP variation developed for this set of experiments was inspired by Stavy and Babai (2008) who developed it as an educational task for schoolchildren. Stavy and Babai designed the perimeter lengths to be either congruent or incongruent with the shape areas. See the left-hand examples in Figure 1. The present study, designed for adults, used both the original stimuli and new stimuli that were more complex in several respects while retaining the misleading nature of the task. For example, one shape in each pair was rotated by 90º relative to the other. See the right-hand examples in Figure 1 and more details below. Notably, the characteristics of this task make it particularly suited for psychometric, math education, and rationality research and practice (e.g., Morsanyi et al., 2018; Stanovich et al., 2016; Toplak et al., 2014). Thus, as required in typical exam contexts, participants were instructed to focus on efficiency (i.e., accuracy plus speed).

Figure 1

Examples of Comparison of Perimeters items.



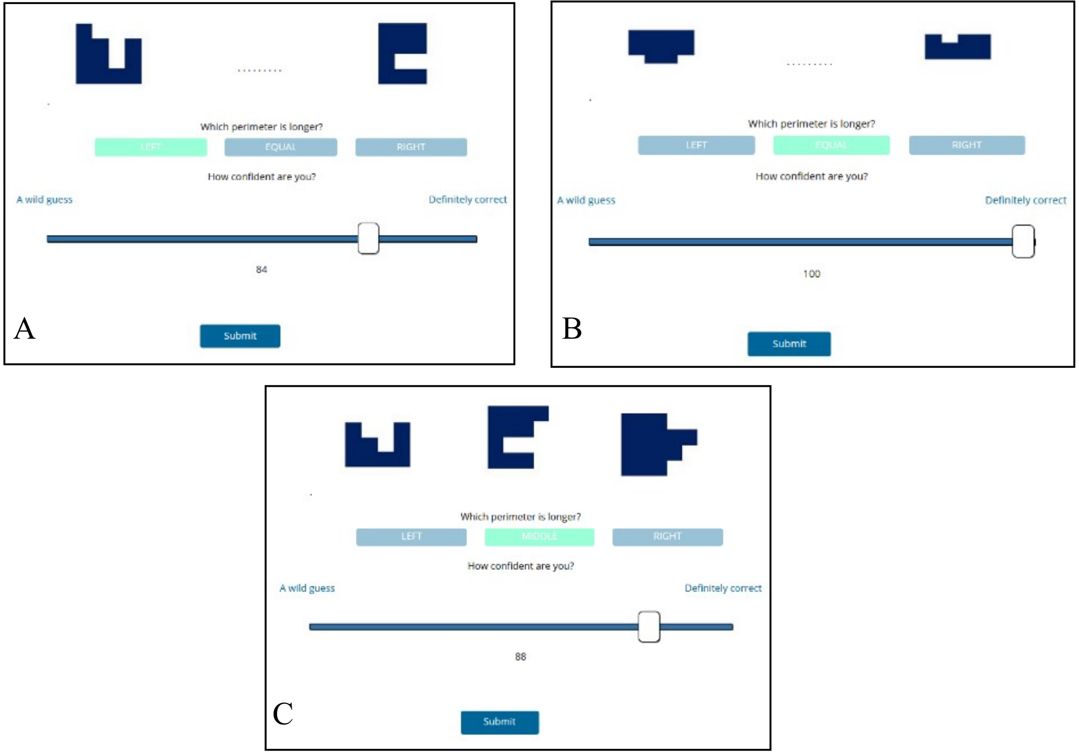
Note. The left-hand (gray) items were used by Stavy and Babai (2008). The percentages are success rates in their study. Congruency exists when the shape with a longer perimeter in the pair also has a larger area. Pairs are considered more complex if both shapes have missing or added squares relative to the 12-square underlying basic rectangle. The right-hand (dark) items were used in the present study, in addition to Stavy and Babai's (2008) items. The dark items may be congruent (top two) or incongruent (bottom two), but all are complex. In addition, in the dark items shown, one of the two shapes in each pair is rotated by 90° relative to the other.

Exp1a was conducted with an online sample of adults from the general public. The study's first aim was to examine the extent of the bias in that population given the incongruency between the shape's perimeter length and its area. Notably, Stavy and Babai (2008) showed the area to be misleading for success, while I focus here on its biasing power for confidence. Based on findings with other misleading tasks (e.g., the "Bat and Ball" famous problem, De Neys et al., 2013), I expected confidence not to properly reflect the challenge generated by area differences when comparing perimeters. A second aim was to identify other cues, either inherent to the task or manipulated, underlying confidence, based on metacognitive theorizing. Identifying these cues has thus far required separate experiments to expose each cue's biasing power, and to demonstrate the difference between its effects on confidence and on success in the task (see Ackerman, 2019, for a review). The BEVoCI method allows for several potential cues to be examined within one task and one sample. Notably, this method allows cues that are not misleading—i.e., that are reflected properly in confidence—to be identified as well. See Figure 2, Panel A for an example of the main task screen.

Exp1b was conducted with a population of advanced undergraduate engineering students. These students have vast experience with challenging, mathematically oriented thinking tasks. One group received a close replication of Exp1a with a 2-shapes comparison. Thus, comparing this group of Exp1b to the general public sample of Exp1a allowed examining cue integration in light of individual differences (level c in Ackerman's [2019] taxonomy). However, critiques regarding commonly used measures of monitoring accuracy have highlighted the effects of performance level on the measurement of resolution and calibration (Fleming & Lau, 2014; Higham & Higham, 2019). Hence, when comparing findings across populations and conditions, difficulty levels should, where possible, be kept at comparable levels, to avoid mistakes in interpreting the results. Towards this end, the stimuli used in Exp1b varied in both height and rotation angle, as detailed below (Figure 2, Panel B).

Figure 2

Screenshots of the Comparison of Perimeters (CoP) tasks used in the present study.



Note. In Panel A the shapes are derived from the 12 squares (in a 3X4 pattern) underlying the basic rectangle, and one of the two shapes is rotated by 90°. The rotation manipulation was used for all stimuli in Exp1a and Exp1c, and for some items in Exp1b. Panel B demonstrates the height manipulation used in Exp1b, in which each square's height was shrunk by 50%, turning them into rectangles. Panel C presents the 3-shapes condition used in Exp1b, with the middle shape in this case having the longest perimeter.

The results of Exp1a indicated, as reported below, that items where the correct answer was EQUAL (as in Figure 2, Panel B)—were more difficult than others. To eliminate this answer option while making minimal changes to the task, the second student group of Exp1b received a variation with three shapes. See Figure 2, Panel C. The 3-shapes version had the same left and right shapes as before, with answer options LEFT, MIDDLE, and RIGHT. When the left and right shapes had equal perimeters, and only in that case, the middle one had a longer perimeter than the others. Thus, the position of the correct response always remained as in the 2-shapes version, with MIDDLE replacing EQUAL. I hypothesized that although the visual load was greater with three shapes than with two, identifying longer perimeters in the central shapes would be easier than identifying equal perimeters. Within Exp1b, comparing two similar task designs—more (2-shapes) and less (3-shapes) misleading—allowed comparing cue integration within the same population.

Exp1c was similar to Exp1a, with another sample from the general public. However, unlike Exp1a, it included detailed instructions intended to improve success rates, by providing visual and verbal explanations of how to systematically count perimeter units, which people are unlikely to apply spontaneously (Galili et al., 2020). The aim was to mimic instructions that accompany new skills being taught or tasks being explained to novices in educational and work contexts, without mentioning any particular misleading cues. One group was instructed to focus on efficiency, as in Exp1a and Exp1b, and the other was instructed to focus specifically on accuracy. The emphasis on accuracy was expected to reduce efficiency but promote success. The detailed instructions and the emphasis on success were also hypothesized to promote depth of processing, which has been associated in the metacognitive literature with better resolution, mainly in reading comprehension tasks (e.g., Dunlosky & Rawson, 2005; Thiede et al., 2003), and with attenuated overconfidence in computerized task performance (Lauterman & Ackerman, 2014; Sidi et al., 2017). As in Exp1b, in Exp1c biases were expected to be attenuated relative to Exp1a here due to the instructions. Given the results of Exp1b (see below), only rotation was used as an additional manipulation, and not height. All these variations were designed to examine whether population and instructional design differences would affect cue validity, cue utilization, and the resultant confidence biases.

The following cues were examined for the CoP task with its variations. The first five cues apply to all groups:

1. **Serial order.** The stimuli were presented in a random order generated for each participant. Typically, in metacognitive research serial order is either ignored or controlled for, as the focus of such research is mostly on item-level processes. Indeed, in the present study as well, controlling for this factor allows us to understand the net effects of other cues. In addition, some findings do point to dissociations between how experience with the task affects success and metacognitive judgments (e.g., Castel, 2008; Kornell & Hausman, 2017; Lauterman & Ackerman, 2014). Examining serial order as a cue can expose both objective and subjective effects of learning from experience, and/or fatigue generated by facing more than fifty items one after the other for about twenty minutes.

2. **Perimeter–area congruency**. Area is the misleading cue suggested by Stavy and Babai (2008), as demonstrated in Figure 1. Congruency with the perimeter is a predictor of success that should be reflected in confidence for it to be reliably utilized. Based on the analyses done by Stavy and Babai, I expected confidence to be insufficiently sensitive to the effect of perimeter–area congruency on success.

3. **Basic shape area.** The basic rectangular shape adapted from Stavy and Babai (2008) was composed of 12 identical small squares arranged in a 3X4 pattern, with squares removed or added on one side only (see Figure 1). In the present study, to increase the challenge, half the items consisted of 24 squares in a 6X4 arrangement, and squares were removed or added on two sides, as part of efforts to increase the task's complexity for adult samples. Differences in the basic underlying shape (12 or 24 squares) were expected to be salient, and thus affect the correspondence between confidence and success. However, the degree of this correspondence was unknown.

4. **Difference in edges.** Shapes with more edges (or corners) than others do not necessarily have longer perimeters. See Figure 1 and Figure 3. In general, the more edges the longer gets the perimeter relative to basic rectangular. However, this is not always the case (see incongruent examples given by Stavy & Babai, 2008 in Figure 1). I expected that the greater the difference in the number of edges between the compared shapes, the easier it would be to identify which shape had a longer perimeter. The question of interest was to what extent confidence reflects this pattern.

5. **Response time**. Time elapsed from when a problem is displayed until participants select their response. This is the most commonly examined cue in metacognitive research, associated with fluency. Overall, both memory and reasoning tasks typically show a negative correlation between response time and metacognitive judgments, reflecting the fact that easy items are processed more quickly than challenging items (e.g., Baars et al., 2020; Koriat et al., 2006; Undorf & Erdfelder, 2015). Ackerman and Zalmanov (2012) demonstrated that response time can be a biasing cue when task items are misleading, with a stronger effect on confidence than on success. From a statistical point of view, unlike the case with other cues, each participant has a different spread of response times across items, with self-paced quick and slow response times. The hierarchical multiple regression models used for the BEVoCI analysis naturally address this statistical complexity. Another aspect

unique to BEVoCI is that including response time in models in addition to other cues exposes its unique contribution above and beyond other cues expected to be associated with item difficulty (or complexity).

Additional cues: As mentioned above, to equate the performance of the general public and engineering students, two manipulated cues were added in conditions in which success was expected to be higher than in Exp1a:

6. **Rotation angle**. Rotation has been examined in metacognitive research with both verbal (e.g., upside-down presentation of words, Sungkhasettee et al., 2011) and non-verbal tasks (e.g., mental rotation, Ariel et al., 2018). In Exp1a, in all pairs, one shape was rotated 90° relative to the other (Figure 2, Panel A). In Exp1b and Exp1c, the values were no rotation (0°), 90°, or 180° (Figure 2, Panel B). Based on findings with the classic mental rotation task (Shepard & Metzler, 1971), I expected that larger rotations would make the task harder (e.g., Parsons, 1995), with potentially a concomitant reduction in confidence.
7. **Base rectangle height**. In Exp1b, with engineering undergraduates, the same stimulus set was used as in the other experiments, but in half the items the height of the component elements was shrunk by half, with the width remaining unchanged. See Figure 2, Panel B relative to Panel A. I hypothesized that this manipulation would make it harder to systematically count the perimeter units, as the manipulated shapes were based on rectangular rather than square units. In reality, the essence of the task remained as before, as both compared shapes were shrunk in the same manner.

All the examined cues were hypothesized to uniquely affect cue validity, cue utilization, or the correspondence between them in at least some of the examined conditions. From a methodological point of view, adding cues to the examined set must be done with care, as cues that have collinearity with others could distort the model results (see Cooksey et al., 1986; Karelaia & Hogarth, 2008). As a rule of thumb, a 0.3 correlation was used as a threshold for collinearity. Three other cues were considered for the present experiment but were excluded due to correlations higher than 0.3 with at least one of the cues defined above. The three excluded cues were related to differences between the two shapes in perimeter length, shape area, and the areas of the imagined rectangles encompassing each shape (maximum extent of additions to the basic rectangle).

2.1. Method

2.1.1. Participants

A power analysis for Exp1a using G*Power (Faul et al., 2009) revealed that for a hierarchical multiple regression model with one sample to achieve a power of 0.90, 48 participants would be required with five cues and 51 with seven cues. Because this experiment was meant to be the basis for the following experiments, which examined two additional cues, a larger sample was used.

For Exp1a, 80 participants were recruited from Prolific Academic (37% females, $M_{\text{age}} = 29.5$, $SD_{\text{age}} = 9.5$). Participants were required to have fluent English and no literacy difficulties based on their self-report to Prolific, an approval rate > 90%, and experience of 50 to 150 tasks (to avoid novices and participants who were too experienced). Participants were asked to devote about 15 minutes to the task without distractions and to use large screens (desktops or tablets) rather than mobile phones. This request was not enforced. The payment was 1.5 GBP, with a promised bonus of 25 pence depending on efficiency.

Inclusion criteria were: (i) Participants had to successfully complete at least three attention verification items out of four (see Materials). (ii) For those who responded particularly quickly (less than 2SD of the sample), the success rate had to be higher than chance (33%). (iii) Confidence had to show some variability unless success was higher than 90% and confidence was 100% for all items. (iv) Focus on the task window had to be higher than 75% of the time spent on the task. (v) A dummy statement included in a self-report after the main task (see Materials) served as another attention check. (vi) At the item level, response time had to be less than 30 seconds (otherwise it was assumed the participant was distracted), and focus time on the window had to be greater than 50% of the time spent on the item. (vii) Finally, participants had to provide at least 45 usable items (out of 54) after the item screening. Participants were excluded only if they violated two or more of these criteria.

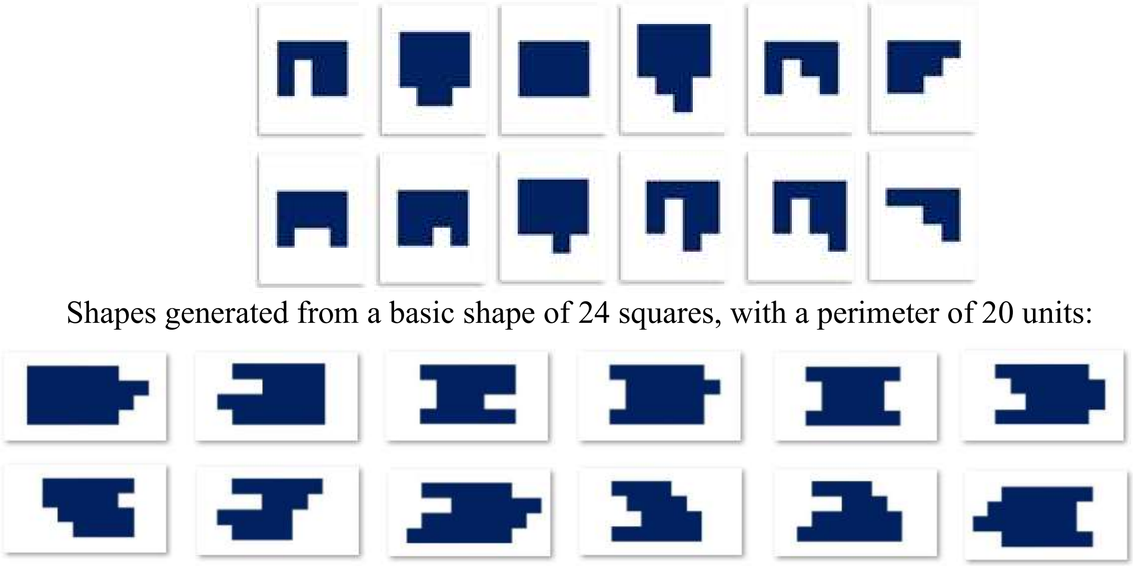
For Exp1b and Exp1c, G*Power indicated that 44 participants would be required in each group to compare two groups via t-tests with an effect size of 0.5 and a power of 0.75. In Exp1b, participants were 105 engineering students (56% females, $M_{\text{age}} = 25.3$, $SD_{\text{age}} = 3.2$). They received a payment of 30 NIS (about 6.5 GBP). In Exp1c, the number of participants in each group was designed to match those of Exp1a. Therefore, 173 Prolific users were recruited (40% females, $M_{\text{age}} = 32.8$, $SD_{\text{age}} = 11.1$). In both experiments, the inclusion criteria were as in Exp1a.

2.1.2. Materials

The stimuli were pairs of shapes with the same basic shape size, 12 or 24 squares. See the entire stimulus pool in Figure 3. Three pairs were used for the instructions and excluded from the data analyses, and four easy pairs were used for attention verification (criterion i above) and were included in the analyses. In addition, for Exp1b, a 3-shapes version of each pair was created, with the new shape added in the middle. The location of the correct solution was unchanged. When the solution was EQUAL in the 2-shapes condition, the middle shape had the longest perimeter, making MIDDLE the correct solution.

Following the main task, participants responded to several self-report questions on a 1–7 scale. The last among them was a dummy statement used as an attention verification criterion (v above); the question instructed participants to ignore the statement and respond with a specific number on the scale.

Figure 3
Stimulus pool used to generate the pairs for the Comparison of Perimeters (CoP) in Experiment 1, in all its variations.



Shapes generated from a basic shape of 12 squares, with a perimeter of 14 units.

Shapes generated from a basic shape of 24 squares, with a perimeter of 20 units:

2.1.3. Procedure

Participants were informed that they would see 54 pairs (or triplets) of geometric shapes based on rectangles. For the 2-shapes comparisons they were instructed to indicate which, if either, of the shapes in each pair, had a longer perimeter, by responding LEFT, EQUAL, or RIGHT. In the 3-shapes comparisons, they were simply asked to identify the shape with the longest perimeter; the EQUAL option was replaced by MIDDLE. A multiple-choice verification question ensured that all participants understood the task before continuing. Participants solved an example and rated their confidence on a continuous scale running from 33%, labeled "A wild guess," to 100%, labeled "Definitely correct." Participants had to move the cursor away from its starting point.

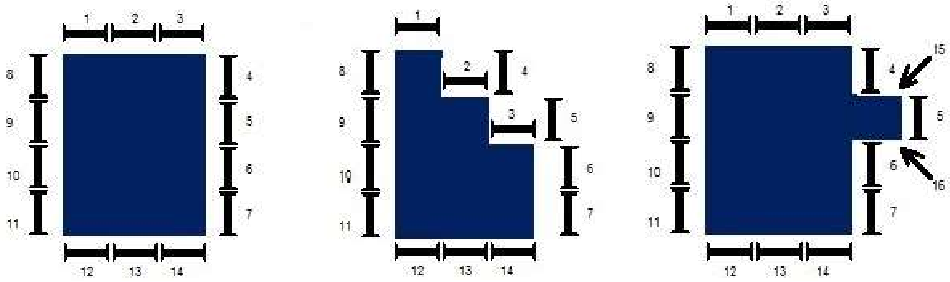
The first two practice pairs were based on 12 squares, while the third introduced a pair based on 24 squares. The third item was quite challenging. The next screen presented its correct answer, which was EQUAL (or MIDDLE), with the statement "This item is certainly not trivial!" Next, the bonus scheme was explained. Participants in all Efficiency conditions were told that previous participants had succeeded in about 70% of the items on average and that their average efficiency was about five correct solutions per minute, with the most efficient participant providing ten correct solutions per minute. Providing at least five correct solutions per minute entitled participants in Exp1a and the Exp1c Efficiency condition to a bonus of 25 pence. In Exp1b, with undergraduates, the expected efficiency was explained, but no monetary bonus was offered. An understanding verification question appeared, followed by a message repeating the emphasis on combining accuracy and speed.

In Exp1b, participants first completed a demographic questionnaire. Demographic information in the other experiments was collected by Prolific Academic.

In Exp1c, with the detailed instructions, participants were shown how to count units along the perimeter of the basic shape (left-hand example in Figure 4). The procedure was then demonstrated with both more (middle example in Figure 4) and less misleading shapes (right-hand example in Figure 4). The misleading nature of the task was not mentioned. For the Accuracy group, it was explained that the average success rate was 70% (based on a pilot study for this condition) and that the most successful participants solved all problems correctly. Those who correctly solved at least 70% of the problems received a bonus.

Figure 4

Illustrations included in the detailed instructions of Exp1c for how to calculate the perimeters of shapes in the Comparison of Perimeters (CoP).



For the task itself, the shape pairs (or triplets) were presented one by one in random order, with the exception of the four attention verification items, which were distributed uniformly for all participants. After every 10 items, a progress message appeared (e.g., “You have so far completed 40 out of 54 items”). The final section comprised the self-report questions with the final attention check.

2.2. Results and discussion

Applying the inclusion criteria for all experiments, in Exp1a, four participants (5%) were excluded, leaving 76 participants. In Exp1b, four participants (4%) were excluded, leaving 101 participants randomly divided into 52 in the 2-shape conditions and 49 in the 3-shape condition. In Exp1c, 15 participants (9%) were excluded, leaving 158 participants randomly divided into 76 in the Efficiency condition and 82 in the Accuracy condition.

2.2.1 Comparing the means among the five groups

Descriptive results for the five groups in the three experiments are presented in Table 1. In the baseline, Exp1a, it is evident that the task, though quick and simple to explain, is not straightforward. Importantly for the purpose of the present study, the success rate of 53% allows room for variability in both success and confidence. Probing into the source for the difficulty revealed that the task was much harder when the correct answer was EQUAL (*M* = 26%) than when it was LEFT (*M* = 64%) or RIGHT (*M* = 84%). This was the basis for designing the 3-shapes condition for Exp1b. As commonly found with many other tasks, overconfidence was pronounced. Response time in this task was similar to that commonly found with memorization tasks (e.g., 4–9 sec., Koriat et al., 2006; 4–7 sec., Undorf & Ackerman, 2017). The main measure that reflects the success–confidence relationship is resolution. The mean resolution in this study was positive, but somewhat lower than that found with memorization of both words and pictures (.28–.45; Bröder & Undorf, 2019; Undorf & Bröder, 2021). Resolution is strongly associated with the misleading nature of the task. A question of interest here is to which cues confidence was sensitive, and to what extent relative to the other considered cues.

Table 1

Experiment 1 – Comparison of Perimeters (CoP): Means (SD) of classical measures in five groups across three sub-experiments.

Experiment + Instruction	Exp1a – Basic	Exp1b – Basic		Exp1c – Detailed	
Population	General public	Undergraduates		General public	
Motivation focus	Efficiency	Efficiency	Efficiency	Efficiency	Accuracy
Task type	2 shapes	2 shapes	3 shapes	2 shapes	2 shapes
Success rate (%)	53 ^{a} (10.3)	63 ^{b} (12.7)	71 ^{c} (16.8)	61 ^{b} (15.8)	66 ^{bc} (18.8)
Confidence (%)	80 ^{ab} (10.1)	78 ^{a} (9.2)	78 ^{a} (8.8)	81 ^{ab} (9.6)	84 ^{b} (10.1)
Overconfidence	26.9 ^{c} (14.4)	15.6 ^{b} (14.5)	7.5 ^{a} (16.4)	20.3 ^{bc} (12.8)	17.8 ^{b} (15.5)
Response time (sec.)	5.0 ^{a} (3.4)	7.9 ^{ab} (4.6)	9.3 ^{b} (6.2)	10.1 ^{b} (7.3)	14.6 ^{c} (10.7)
Efficiency (correct answers/min.)	8.2 ^{d} (3.6)	6.3 ^{bc} (3.3)	7.1 ^{cd} (5.0)	5.1 ^{ab} (2.8)	4.2 ^{a} (2.7)
Resolution (gamma)	.22 (.21)	.26 (.20)	.34 (.30)	.24 (.25)	.24 (.34)

^{a,b,c} Significant pairwise differences in dependent variables between groups across all experiments, $p \leq .05$.

The 2-shapes condition in Exp1b was highly similar in its methodology to Exp1a, with a different population and two additional manipulated cues. A one-way Analysis of Variance (ANOVA) comparing the two experiments pointed to a higher success rate in Exp1b, despite the two additional cues, and equivalent confidence, which resulted in attenuated overconfidence. More time was invested by Exp1b's undergraduates, suggesting higher motivation, despite the absence of a monetary bonus. However, this group showed lower solving efficiency than the sample from the general public in Exp1a, reflecting that although more time was invested, the success rate did not rise proportionally. Thus, we see differences between Exp1a and its close replication. Notably, the resolution was equivalent to that found in Exp1a, which is central to using BEVoCI. The question of interest is whether this similar resolution "hides" differences in the cues predicting success and confidence.

Comparing the 3-shape condition to the 2-shape conditions of Exp1a and Exp1b, given the highly similar tasks and response requirements, the differences are striking. First, as expected, success in the 3-shape condition was greater than in both 2-shape conditions. This finding helps reject overload as a reason for the changes in cue integration. Mean response time was longer than in Exp1a, and efficiency was between the efficiency levels seen in the previous 2-shapes conditions, but not significantly different from either. Mean confidence ratings were blind to the higher success rates in the 3-shapes condition, yielding attenuated overconfidence.

Turning to Exp1c, as can be seen in Table 1, almost no differences were found between the Efficiency and Accuracy conditions in the classic dependent variables, but in both cases, success rates were higher than in Exp1a. This came at the price of longer response times in the Accuracy group than in the Efficiency group; and in both cases, time invested was more than double the time participants invested in Exp1a, which harmed efficiency. Confidence and resolution were not affected by the instructions, while overconfidence was significantly attenuated only in the Accuracy group.

2.2.2 BEVoCI—Cue utilization and cue integration across the five groups

Now, we turn to the BEVoCI methodology to consider the relative contribution of each cue and to compare cue validity and cue utilization. The code and data appear in OSF: https://osf.io/jgn54/?view_only=d66bed9658d14932ab8a1bd86b1dec97. As with any within-participant correlations, gamma used for calculating resolution included, the first stage was to exclude participants who showed no variability in either confidence or accuracy (i.e., measures for all items were the same). These included two participants (2%) in Exp1b, four (2.5%) in Exp1c, and none in Exp1a.

As described above, a methodological advantage of BEVoCI is that multiple cues can be examined at the same time, while controlling for all other cues. Prior to the analyses, all independent and dependent variables were standardized. All cues were entered into one hierarchical multiple regression model predicting success and another predicting confidence for each experimental condition separately using the R 4.1.0 package nlme

(Pinheiro et al., 2019).[1] In both models level 1 was the items and level 2 was the participants. Data were modeled as a linear function of the cues related to each response, plus the participant's "random" shift. The beta coefficients and their significance are reported in Table 2.

[1] R command for Experiment 1: model.success<-lme(Accuracy_dv_c ~ 1 + Serial_order_iv_c + Perimeter_area_congruency_iv_c + Basic_shape_area_iv_c + Difference_in_edges_iv_c+ Response_time_iv_c, random = ~1|Username, data = raw_data). All predictors were centralized (denoted by c). The model for predicting confidence was the same with Confidence_dv_c instead of Accuracy_dv_c.

Table 2

Cue validity and cue utilization in Experiment 1 with the Comparison of Perimeters (CoP): Results of hierarchical multiple regression analyses, presented as standardized β of all relevant cues for each experimental group when predicting success and confidence.

Experiment	Exp1a – Basic		Exp1b – Basic + Rotation + Height				Exp1c – Detailed instructions			
Population	General public		Undergraduates				General public			
Manipulation	Efficiency		Efficiency		Efficiency		**Efficiency**		**Accuracy**	
	2 shapes		**2 shapes**		**3 shapes**		2 shapes		2 shapes	
DV IV	Success	Confidence	Success	Confidence	Success	Confidence	Success	Confidence	Success	Confidence
Serial order	.01	-.01	.06**	-.03	.02	.03	.00	-.04**	-.01	-.06**
Perimeter–area congruency	.41***	.01	.20***	.01	.11***	.02	.31***	.03**	.28***	.00
Basic shape area	-.08***	-.19***	-.14***	-.20***	.03	-.15***	-.08***	-.16***	-.07***	-.17***
Difference in edges	.20***	.12***	.20***	.12***	.07***	.07***	.03*	.07***	.02	.04***
Rotation angle			-.04	.00	-.04*	.00	.01	.02	.00	-.02
Height			.05**	.04**	.06**	.04*				
Response time	.02	-.07***	.03	-.07***	.03	-.13***	.02	.00	.03	-.03

Note. Significance of a cue as a predictor, *** $p \leq .001$; **$p \leq .01$; *$p \leq .05$

Gray font: A match between cue validity and cue utilization. **Bold** fonts: Significant mismatch between the association of the cue with success and with confidence, $p < .05$. **Larger** font indicates a change in direction, from no significant association to a significant one, or opposite associations of the cue with success and with confidence.

In the next stage, a combined model was used. In this model the data were duplicated, once with Measure_type = “success” and Measure taking success (dichotomy), and once with Measure_type = “confidence” and Measure taking confidence (continuous). This allowed comparing the contribution of each cue to the two measures by obtaining a significance level for the cue*Measure_type interaction. The differences are highlighted in Table 2. The aim was to expose the cues that generate biases, through confidence being over- or undersensitive to changes in success.

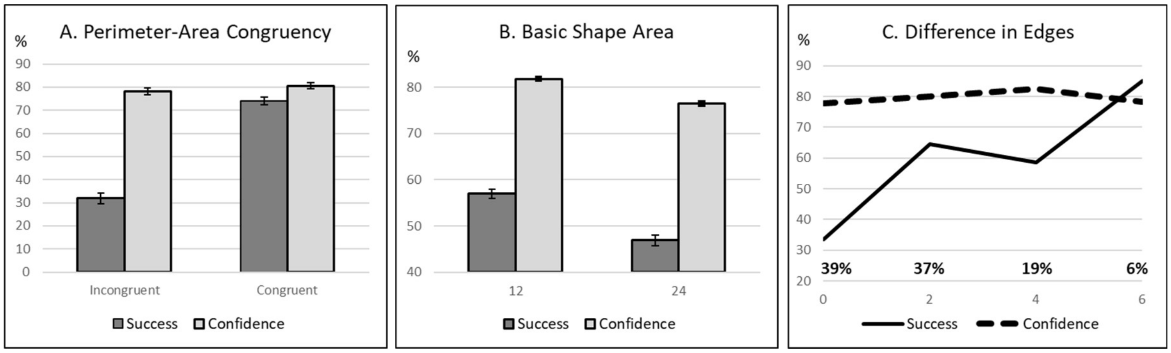
Starting with the baseline in Exp1a, the comparison between cue validity (predicting success) and cue utilization (predicting confidence) is striking. Perimeter–area congruency, the misleading cue identified by Stavy and Babai (2008) with much easier task items, was a strong predictor of success here as well. Notably, however, it was not used as a cue for confidence despite the stimuli being more complex than the original set, and there was no ceiling effect. Figure 5 Panel A clearly displays the insensitivity of confidence to the perimeter–area congruency, with overconfidence in the incongruent task items but good calibration in the congruent task items. This good calibration suggests that the task itself was not misleading, but that area had a particularly misleading power.

Basic shape size (12 or 24 squares) negatively affected success in the task, and confidence reflected this direction properly with a slight bias. See Figure 5 Panel B. Nevertheless, the BEVoCI analysis showed bias in this cue as well, with the regression coefficients for confidence significantly stronger than for success. This bias is not immediately evident in the figure, though it should be noted that figures do not exclude variance explained by other cues in the regression models. The difference in predictive power may originate in the smaller errors of the mean for confidence than for success. This is a type of bias that so far has been overlooked in metacognitive research.

In Figure 5 Panel C, the bias generated by the difference in edges is clearly evident, as confidence was less affected by this cue than success. In particular, when the difference in edges was small, participants showed considerable overconfidence, while when this difference was large their confidence was well-calibrated. Even after controlling for all other cues associated with the perception of difficulty, response time still generated a bias, though weak. The finding that serial order did not affect either success or confidence is notable as well. It is possible that learning from experience and fatigue balanced each other out. Confidence, though, was not biased by serial order.

Figure 5

Exp1a—Examples of biases in the Basic Efficiency condition.



Note. Panel A demonstrates insensitivity of confidence to changes in success due to perimeter–area congruency. Panel B demonstrates oversensitivity of confidence to changes in success, expressed mainly in attenuated variance. In both panels, error bars represent standard errors of the means. Panel C demonstrates insufficient sensitivity of confidence ratings to the effect of the difference in edges. The percentages above the X-axis show the proportion of pairs characterized by an edge difference of 0, 2, 4, or 6 units.

As can be seen, Exp1a provided an initial probe into sources of bias in a novel meta-reasoning task. Notably, the results already demonstrate the value of BEVoCI in considering multiple cues simultaneously. Furthermore, in using a visual task, the experiment exposed heuristic cues and biases not considered previously. In particular, the results revealed four different types of dissociations between success rate and confidence, showing a decisive double dissociation between cues that predict success and those predicting confidence. First, for perimeter–area congruency, confidence was totally blind to effects on success; second, for the difference in edges, confidence was insufficiently sensitive to the cue; third, for basic shape area, confidence was too sensitive to cue variations; and fourth, for response time, confidence showed a false negative tendency. All these effects were found within the same task and group of participants. Thus, clearly, different factors affect success and confidence and call for attention as part of the effort to identify biases and ways to overcome them.

Turning to comparisons among the conditions, I first describe the BEVoCI findings from Exp1b's 2-shapes conditions, which replicated Exp1a with the undergraduate population. See Table 1 and Table 2. As expected, the engineering undergraduates achieved better success, but interestingly, their resolution remained comparable, meaning the engineering students were no better than the general public at discriminating between correct and wrong responses, despite their greater experience with mathematically oriented challenges. All cues which affected either success or confidence in Exp1a—perimeter–area congruency, basic shape area, difference in edges, and response time—generalized with this population. Several differences were also found. Serial order was positively predictive of success, indicating learning from experience here but not in Exp1a. However, this learning was not reflected in confidence ratings and thus generated a bias. Perimeter–area congruency was still misleading, but its effect on success was attenuated relative to Exp1a, and it was still not reflected in confidence. Concerning the two new cues, the rotation angle was not predictive of success and was not misleading for confidence, unlike what was expected based on the mental rotation task (Parsons, 1995). The height manipulation predicted success and was reflected properly in confidence.

The malleability of cue weights is evident in the 3-shapes version of Exp1b. Four cues were utilized properly for confidence: serial order, number of edges, rotation angle, and height. To statistically compare the two versions of the biases that remained, I ran two regression models, similar to the BEVoCI model, one for success and one for confidence, with the addition of the task version and its interaction with each cue. The biasing effect of the perimeter–area congruency on success was further attenuated relative to the 2-shapes version, $t(5294) = 3.82$, $p < .0001$ (the latter was already attenuated relative to Exp1a). However, perimeter–area congruency was still ignored in confidence ratings, with no difference between the two task versions, $t(5294) = .28$, $p = .78$. Basic shape area became non-predictive of success, which is impressive and significantly different from the findings with the 2-shapes version, $t(5294) = 6.12$, $p < .0001$; but its effect on confidence remained, though somewhat attenuated, $t(5294) = 1.76$, $p = .08$, indicating a more pronounced bias (larger beta difference) than in the 2-shapes version. The biasing effect of response time also grew, as there was no change in its association with success, $t(5294) = .62$, $p = .53$; but it became more strongly associated with confidence, $t(5294) = 2.69$, $p = .007$. Thus, the 3-shapes version was easier in terms of success and was also overall less misleading than the 2-shapes version of the task, but the detailed analysis did expose biases that were weaker in the 2-shapes version.

Generalizing Exp1a, both conditions in Exp1b revealed double dissociations, with cues affecting success more than confidence, confidence more than success, neither, or both. Thus, although the engineering undergraduates did better in the task, they were misled by item-level cues in a way comparable to the general public.

Exp1c included detailed instructions for the task in two conditions, with an emphasis either on efficiency, as in the previous conditions, or on accuracy. As mentioned above, success improved with the detailed instructions, but at the price of efficiency. As for cue validity and cue utilization, looking into Table 2 reveals that the detailed instructions were effective in both groups in eliminating or attenuating most biases found in Exp1a. However, both perimeter–area congruency and basic shape area remained persistently predictive of success and misleading for confidence, showing the double dissociation of confidence as insufficiently sensitive to the former and oversensitive to the latter. Comparing the two groups of this experiment statistically revealed few significant differences, with none having $p < .045$ (details available on request). Minor biases were generated by serial order and response time. On the other hand, clarifying the inconsistent results in Exp1b, rotation angle was non-predictive of success and not biasing for confidence.

To sum up, as suspected by Stavy and Babai (2008), the CoP is a highly misleading task that generates both response mistakes and confidence biases. The present experiment demonstrates that this is the case not only for schoolchildren but also for adults—even adults with strong mathematical backgrounds or who received detailed instructions. However, the focus of the present study goes beyond this particular task. Using BEVoCI allowed exposing impressive cue integration and its malleability across populations, instructions, and task variations, despite having similarly mediocre resolutions across them. This study demonstrates the effectiveness of exposing multiple sources for mistakes and confidence biases within one group of participants. In terms of educational design, BEVoCI allows exposing which bias sources are stronger than others, which were attenuated by instructional manipulations, and which remained persistent even when success rates improved. In particular, we see that background knowledge (Exp1b) did not immunize participants against confidence biases, while detailed guidance with instructions designed to promote efficiency attenuated most biases (Exp1c).

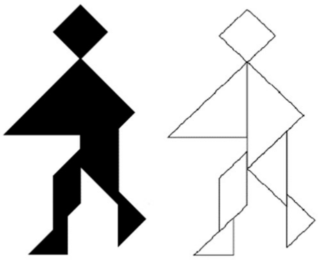
3. Experiment 2: Exposing Biases in a Novel Task

Experiment 1 used a task where prior knowledge was available about a central factor, namely perimeter–area congruency, likely to be misleading for confidence. Experiment 2 demonstrates the use of BEVoCI when no prior knowledge about biasing factors for confidence is available.

The *Missing Tan Task* (MTT) is a novel non-verbal task. It is based on a physical game called Tangram (https://en.wikipedia.org/wiki/Tangram; see Figure 6), in which silhouettes must be formed by positioning seven geometric pieces (*tans*)—a square, a parallelogram, two large triangles, two small triangles, and one intermediate triangle. As mentioned above, this game has been used in educational settings to help teach geometry and spatial reasoning (Bohning & Althouse, 1997; Lee et al., 2009). In the multiple-choice task developed for the present study, all silhouettes were generated from six out of the seven pieces (see Figure 7). Participants' task was to identify the missing piece when viewing the static silhouette, without manipulating physical pieces.

Figure 6

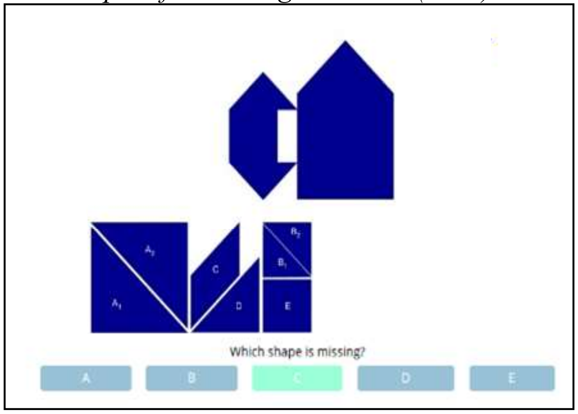
An example of the original Tangram game.



Note. Players are presented with a silhouette (left) and must position seven geometric pieces (*tans*) to form the solution (right).

Figure 7

An example of a Missing Tan Task (MTT) item.



Which shape is missing?

Pilot studies were used to develop a stimulus set comparable to the basic version of the CoP in global task difficulty and item difficulty range. Since MTT items take longer to solve than CoP items, only 30 items were used rather than 54—still enough to support robust within-participant analyses. The instructions, bonus schemes, and online sample selection criteria were as similar as possible to those in Experiment 1.

In the absence of prior research with this task, general metacognitive research as well as insights from the CoP offered potential biasing factors. The following cues were hypothesized to affect objective and/or subjective difficulty:

1. **Serial order.** As before, experience with the task may yield benefits from learning, detriments due to boredom and fatigue, or both.
2. **Silhouette area.** Unlike in the CoP, in the MTT, the area can be a relevant cue for success. Silhouette area can take three values, depending on the missing piece: one of the two large triangles (A in Figure 7); one of the two small triangles (B); or one of the three other pieces (C, D, and E), which differ in shape but have the same area. The example in Figure 7 is of the last type because the missing piece is the parallelogram (C). If a triangle A or B is missing, the silhouette area uniquely indicates the correct answer, which might make the task easier. However, smaller shapes have more potential locations in the silhouette than larger ones. This consideration leads to the prediction that the task is hardest when B is missing. The question is whether confidence properly reflects the factors that indeed affect success.
3. **Rotated pieces.** The number of pieces that must be rotated or flipped relative to their representation in the legend. Although when using the CoP, rotation did not affect success or confidence, classic studies indicate that tasks which require mental rotation are more challenging (Shepard & Metzler, 1971). In the example in Figure 7, the two large triangles (A) and the square (E) are oriented as in the legend, while the other three pieces, both B pieces and D, must be rotated.
4. **Perceived nameability.** Lauterman and Ackerman (under review) found the perceived nameability of components within Raven matrices to be a misleading cue for metacognitive judgments of solvability. Here, in a pre-test with a different sample (*N* = 80), participants judged on a Likert scale how easy it was to name each silhouette. The Tangram and MTT tasks were not mentioned. The means were attached to each stimulus.
5. **Description length.** The same pre-test sample was asked to describe each silhouette shape as briefly as possible. Shorter descriptions were expected to represent more familiar shapes than those requiring more words. Familiarity was expected to be a misleading cue that could distract participants from the components of the silhouette,

similarly to the misleading effect of familiarity and accessibility underlying other metacognitive judgments (Ackerman & Beller, 2017; Koriat & Levy-Sadot, 2001). Thus, the average number of words used in the description was used as a potential cue.

6. **Response time.** Time elapsed from figure display until participants chose one of the answer options, as in Experiment 1.

The number of edges in each silhouette, the length of the perimeters, and the areas of the imagined rectangles encompassing each silhouette were also considered here as potential cues. These attributes had correlations greater than .30 with other considered cues and thus were not included in the analyses.

As in Experiment 1, I aimed to identify which considered stimuli characteristics predict success, and then to identify underutilization, overutilization, and well-adjusted cues for confidence. Exposing biasing factors is required for guiding future attempts to improve solving success and monitoring accuracy.

3.1. Method

3.1.1. Participants

The sample size was chosen to be similar to that used in the basic condition of the CoP with the general public (Exp1a). The initial sample comprised 85 Prolific users (41% females, M_{age} = 26.5, SD_{age} = 7.2). The basic payment was 2.4 GBP with a potential bonus of 25 pence.

3.1.2. Materials

The main stimuli were thirty silhouettes, each comprised of six out of the seven possible pieces. The legend showing the seven pieces and the question “Which shape is missing?” appeared on the screen throughout the task. See Figure 7. The confidence scale ranged between 20% (chance level) and 100%.

3.1.3. Procedure

The procedure was as similar as possible to that of Experiment 1. In particular, participants were incentivized to focus on efficiency. At the start of the task, participants solved three practice problems, which did not count toward the bonus.

3.2. Results and discussion

Four participants were excluded based on the same selection criteria as in Experiment 1, leaving 81 participants for the data analyses. Descriptive results are presented in Table 3, in the same format as in Table 1. Comparing the findings to the basic condition of the CoP (Exp1a), participants achieved similar success rates. This was an important goal when designing the tasks. Items in the MTT task required more time than solving CoP items. The average time invested was similar to that spent on intermediate-difficulty Raven matrices (Ackerman et al., 2020).

Before the main analyses, participants were screened for lack of variance in success or confidence, as in Experiment 1. No one had to be excluded under this criterion.

Table 3
Experiment 2 – Missing Tan Task: Means (SD) of classic measures.

Instructions	Basic
Population	General public
Motivation focus	Efficiency
Success rate (%)	49 (14.4)
Confidence (%)	73 (8.1)
Overconfidence	23.6 (15.8)
Response time (sec.)	22.4 (9.6)
Efficiency (corrects/min.)	1.6 (0.8)
Resolution (Gamma)	.41 (.29)

Table 4
Cue validity and cue utilization in Experiment 2 – Missing Tan Task (MTT) with the general public: Results of hierarchical multiple regression analyses, presented as standardized β of all relevant cues when predicting success and confidence.

DV	Success	Confidence
IV		
Serial order	**.06****	**-.02**
Silhouette area	**-.09*****	**.01**
Rotated pieces	.16***	.10***
Perceived nameability	**-.07*****	**.03**
Description length	.05**	.01
Response time	**-.08*****	**-.27*****

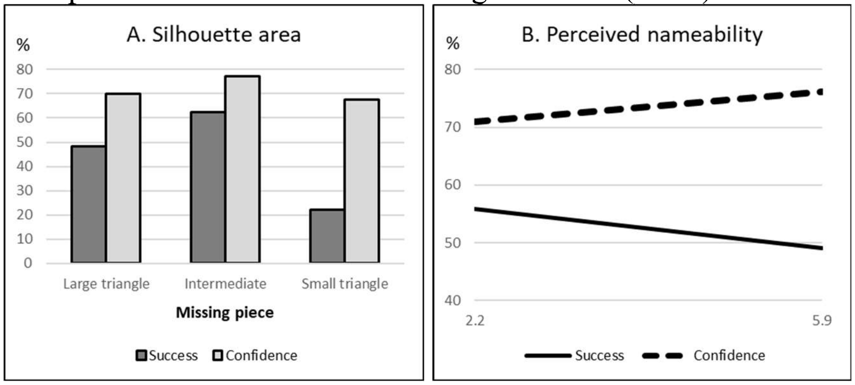
NOTE. Significance of a cue as a predictor, *** $p \leq .001$; **$p \leq .01$; *$p \leq .05$
Gray font: A match between cue validity and cue utilization. **Bold** fonts: Significant mismatch between associations of the cue with success and with confidence, $p < .05$. **Larger** font indicates a change in direction, from no significant association to a significant one, or opposite associations of the cue with success and with confidence.

The beta coefficients of the BEVoCI models are presented in Table 4. All six examined cues were predictive of success, while only two reliably predicted confidence. Specifically, the number of rotated pieces and description length were utilized properly, while there was overutilization of response time and underutilization of serial order, silhouette area, and perceived nameability. Thus, we see here again a clear double dissociation between effects on success and on confidence.

Figure 8 presents two discovered biases. First, the results support the possibility that the task was hardest when one of the smallest triangles was missing, on the basis that the smaller the piece, the greater the number of places where it could possibly be located in the silhouette. Confidence was not sensitive enough to this challenge, which resulted in pronounced overconfidence.

Figure 8

Examples of two biases in the Missing Tan Task (MTT).



Note. Panel A exposes that the task is hardest when a small triangle is the missing piece, but this is overlooked by participants, generating pronounced overconfidence. Panel B shows that perceived nameability affects success and confidence in opposite directions.

Second, the associations of perceived nameability with success and confidence were in opposite directions, generating a bias. This bias source resembles other misleading cues associated with processing fluency (e.g., Ackerman et al., 2013; Koriat, 2018; Thompson et al., 2013), but here we see this pattern with a continuum, rather than discrete levels, as seen with other cues previously examined (e.g., font size, Undorf & Zimdahl, 2019).

To sum up, applying the BEVoCI data analysis to a novel task is clearly effective in exposing biases. The cues considered here were less predictive of confidence than in the CoP, as more variance remained to be explained by response time. This finding suggests that future research may consider additional potential biasing sources for the MTT. More broadly, this experiment provides food for thought toward intervention attempts. Pinpointing pitfalls as demonstrated here (e.g., overconfidence was greatest when the smallest shape was missing) has broad implications for educational design, as these are the points learners need help with, and where they have the greatest potential for improving their success. See the General Discussion.

4. Experiment 3: Reanalyzing Existing Data from an Educational Context

A major advantage of BEVoCI is that it allows exposing biases in existing data. While this facility cannot replace systematic manipulations that control for alternative explanations, from a practical perspective the ability to expose several biases at once using data already collected benefits both research and practice.

To demonstrate this strength of BEVoCI, I searched for published papers that met the following criteria: tasks with a clear-cut correct response for each task item, making it possible to decisively distinguish correct from wrong responses; several tens of items for each participant; more than a hundred participants; a metacognitive judgment collected for each item; and the presence of potentially misleading cues inherent within item characteristics, whether already recognized within the data or identifiable based on metacognitive theory. For coherence, I also aimed to find a study done in an educational context, using ecologically valid tasks that were similar in nature to those used in the other experiments of the present study. However, these latter specifications are not essential for using the BEVoCI method to expose cues and metacognitive biases.

I chose to reanalyze data published online as part of research by Vuorre and Metcalfe (2022). The study included several tasks taken from the Regents exams administered by the Office of State Assessment of New York. I focused specifically on participants who solved algebra questions (https://www.nysedregents.org/algebraone/). These were included in two experiments with eighth-graders (1A and 1B, *N* = 84 in the algebra condition) and two experiments with adults (1E, *N* = 86, and 1F, *N* = 92). Vuorre and Metcalfe's experiment 1B was a direct replication of their 1A, and so the two were merged in the present analysis. To avoid confusion with the experiments of the present study, hereafter I refer to their experiments as VM1AB, VM1E, and VM1F.

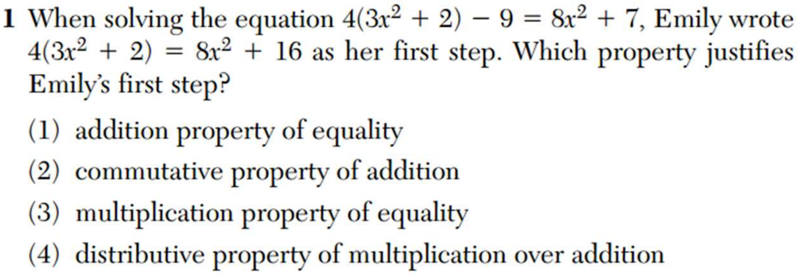
All experiments made use of the same set of problems. Each participant solved 40 problems, in either two (VM1E) or four (VM1AB and VM1F) sessions. The dataset contained participants with missing data. As I intended to use session number as a cue in the BEVoCI analysis, I included only participants with data for at least 25 problems (meaning that the data must have been collected in more than one session). VM1AB and VM1E were done on paper, with confidence scales of 0–5 for VM1AB and 0–10 for VM1E. VM1F was computerized, with a 0–100 confidence scale. For consistency, the 0–5 and 0–10 values were converted to a 0–100 scale by multiplying by 20 and 10, respectively. However, the fact that the original scales were not in percentage terms means that they cannot be used to assess over- or underconfidence (see Ackerman, 2019).

Six cues were considered in the reanalysis as potentially biasing for confidence. See Figure 9 for an example. The cues were:

1. **Session number.** VM1AB took place over four sessions on separate days, with 10 problems in each. The children received a feedback session between tests. VM1E, with undergraduates, took place in two sessions held on the same day, with 20 problems in each; participants received a mathematical tutorial after the first. VM1F, also with undergraduates, involved four tests in one meeting, with 10 problems in each. Thus, session number was available for VM1AB and VM1E, but not for VM1F.

Figure 9

Example of an algebra problem used in Vuorre and Metcalfe (2022).



1 When solving the equation $4(3x^2 + 2) - 9 = 8x^2 + 7$, Emily wrote $4(3x^2 + 2) = 8x^2 + 16$ as her first step. Which property justifies Emily's first step?

(1) addition property of equality

(2) commutative property of addition

(3) multiplication property of equality

(4) distributive property of multiplication over addition

Note. A problem taken from the June 2014 Regents exam, with the identifier A-REI.A. The problem text included three lines, no graph, and a formula.

2. **and 3. Problem characteristics.** Each problem had a non-unique identifier provided by the Regents board, which reflects problem characteristics relevant for state assessment purposes (e.g., A-REI.C). The first letter, A, indicates an algebra problem and appeared in all problems within the included data. The next three letters represent

four problem types: APR (Arithmetic with Polynomials and Rational Expressions); REI (Reasoning with Equations and Inequalities); CED (Creating Equations); and SSE (Seeing Structure in Expressions). The last letter had four values as well, A, B, C, and D[2]. I included the middle and last components as potential cues, recoding them based on the proportion of each type that was answered correctly across all participants. This recoding by success rate allowed examining whether confidence follows the same order as success rates for each characteristic. My numerical codes for the middle component were 1=APR, 2=REI, 3=CED, and 4=SSE, and for the last component they were 1=B, 2=C, 3=A, and 4=D.

Based on previous research, I categorized all problems by three other potential misleading cues. See Figure 9.

4. **Question text length**. Based on cognitive load theory, instructional text length is negatively related to performance (e.g., Leahy & Sweller, 2016; Walkington et al., 2015). I categorized question texts into two groups: those with few lines of text (1 or 2, 50% of the problems) and those with many (3 or more). I tested whether text length is indeed associated with success, then examined the extent to which this association is reflected in confidence.
5. **Graphs (yes/no)**. Classic guidelines for teaching mathematical concepts include visual representation as a key to designing successful study materials (Rau & Matthews, 2017; Tindall-Ford et al., 2020). However, in metacognitive research, several studies have pointed to potential upwards bias in metacognitive judgments generated by the concreteness of visual representations (Ackerman & Leiser, 2014; Ackerman et al., 2013; Serra & Dunlosky, 2010). In the dataset, about 14% of the problems included a graph. Assuming that the graphs were chosen to be helpful, the question is whether confidence reliably reflects the help they provide.
6. **Formulas (yes/no)**. A central challenge in algebra education is the interpretation of symbolic representations (Capraro & Joffrion, 2006). In the dataset, 69% of the

[2] More details about the coding scheme can be found at https://www.jmap.org/JMAPArchives/CurrentVersion/JMAPAI_REGENTS_BOOK_BY_PI_TOPIC.pdf.

problems included a formula. I examined how the presence of a formula affected success and confidence.

Using a BEVoCI analysis enables addressing the following research questions: (i) What considered cues indeed predict success in this task? (ii) What cues underlie confidence? (iii) What is the relative balance among the considered cues? (iv) Which cues bias confidence relative to their predictive value for success? (v) In what respects are the three groups similar and different in cue validity and cue utilization? The group comparisons are of particular interest given previous findings that metacognitive processes when facing complex tasks are almost mature by the eighth grade (Koriat et al., 2014), and in light of screen inferiority relative to working on paper in both cognitive and metacognitive processes (lower performance, larger overconfidence, and less effective adjustment to task conditions) in problem-solving tasks (Sidi et al., 2016; Sidi et al., 2017) and in reading comprehension tasks among children and adults (Delgado et al., 2018; Golan et al., 2018; Lauterman & Ackerman, 2014). Thus, this reanalysis can inform both developmental and human-computer interaction research.

4.1. Method

4.1.1. Participants

The analyzed dataset included 262 participants: 84 middle-school children in VM1AB, 86 undergraduates in VM1E, and 92 undergraduates in VM1F.

4.1.2. Materials

The materials were 42 algebra problems from the Regents high school examinations administered in 2014 and 2015 in a multiple-choice test format. See example in Figure 9.

4.1.3. Procedure

The relevant procedure characteristics were detailed above. In particular, they included running the experiment on paper (VM1AB, VM1E) vs. on screen (VM1F); with confidence scales of 0–5 (VM1AB), 0–10 (VM1E), or 0–100 (VM1F); and with two (VM1E) or four (VM1AB, VM1F) sessions, taking place on the same day (VM1E, VM1F) or on separate days (VM1AB). The task included 40 items per participant in all designs.

4.2. Results and discussion

Descriptive results of the three groups are presented in Table 5, as for Experiment 1.

Table 5

Experiment 3 – Reanalyzed algebra data from Vuorre and Metcalfe (2022): Means (SD) of classic measures.

Experiment	VM1AB	VM1E	VM1F
Population	8th graders	Undergrads	Undergrads
Medium	Paper	Paper	Screen
Success rate (%)	69.3 (15.5)	71.3 (19.4)	73.7 (18.0)
Confidence (transformed to 0–100)	71.3 (17.0)	69.1 (21.7)	71.1 (21.0)
Overconfidence	---	---	2.0 (13.2)
Resolution (gamma)	.52 (.38)	.54 (.43)	.45 (.46)

Response time was not available in the dataset, and thus efficiency could not be calculated. The transformed confidence ratings (VM1AB, VM1E) reached levels highly similar to confidence elicited by a percentage scale. Nevertheless, as mentioned above, overconfidence cannot be calculated for VM1AB and VM1E, where the confidence scales were not in percentage terms. Interestingly, there were no significant differences among the groups, all *p*s > .24. Mean success rates and resolution were higher than in the previous experiments, though neither reached the point of risking ceiling effects, while confidence was similar to that previously found.

The BEVoCI analyses begin by examining collinearity among the cues considered. Given that session number was available in VM1AB and VM1E but not VM1F, its correlation with the other cues was calculated first. No correlation passed the .30 threshold. Thus, session number is used as a cue for VM1AB and VM1E.

Examining correlations among the other cues across all three groups revealed that the middle component of the problem identifying code (the three letters) was correlated at .36 with the presence of formulas, and at -.31 with the presence of graphs. This correlation likely reflects the propensity for certain problem types to contain formulas or graphs. Therefore, this cue was omitted from further analyses. The presence of graphs and formulas was negatively correlated, -.34. Although most problems included either a formula or a graph, there were nine problems (21.4%) containing neither and two problems (4.8%) containing both. Thus, they could not be combined into one variable. I chose to keep formulas as a cue over graphs because 69.0% of the problems had formulas and only 14.3% had graphs, making the data split between problems with and without formulas more balanced and powered than for graphs. If there were theoretical reasons guiding favoring having graphs as a cue over formulas, this could be done as well. Thus, four cues were included in the present BEVoCI analyses: session number for VM1AB and VM1E, but not VM1F; and for all experiments, the last letter of the Regents standard identifier (A–D), text length (few/many lines), and presence of a formula (yes/no).

Prior to the main analyses, participants were screened for variance in success rates and confidence ratings. This screening excluded seven participants who showed no variance in at least one measure (2.7%).

Table 6

Cue validity and cue utilization in Experiment 3, reanalyzed algebra data from Vuorre and Metcalfe (2022): Results of hierarchical multiple regression analyses, presented as standardized β of all relevant cues when predicting success and confidence.

Experiment	VM1AB		VM1E		VM1F	
Population	8th graders		Undergraduates		Undergraduates	
Medium	Paper		Paper		Screen	
DV	Success	Confidence	Success	Confidence	Success	Confidence
IV						
Session no.	-.03	-.06***	.07*	.06*	---	---
Last letter	.08***	.09***	.08***	.10***	.10***	.11***
Text length	-.07*** [a]	-.10***	-.06*** [ab]	-.08***	**-.02** [b]	**-.10*** **
Formula	**.10*** **	**0**	**.03**	**-.02**	**.03**	**-.04** **

NOTE. Significance of a cue as a predictor, *** *p* ≤ .001; ***p* ≤ .01; **p* ≤ .05
Gray font: A match between cue validity and cue utilization. **Bold** fonts: Significant mismatch between associations of the cue with success and with confidence, *p* < .05. **Larger** font indicates a change in direction, from no significant association to a significant one, or opposite associations of the cue with success and with confidence.

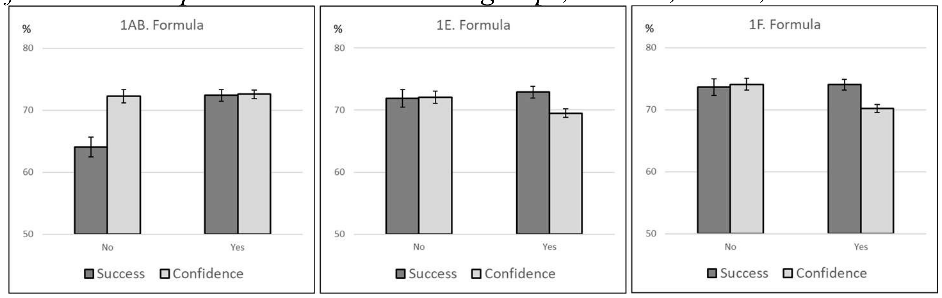
Table 6 presents the BEVoCI results. Clearly, all considered cues had predictive power for success and for confidence in at least some conditions. The analysis of session numbers for the VM1AB group showed no improvement with learning despite receiving feedback between sessions, and confidence was even reduced. The adults in VM1E improved from one session to the next, and this was reflected properly in confidence. Comparing the VM1AB and VM1E groups revealed that the adults benefited significantly more from experience, and this was reflected in both success, *t*(6355) = 2.84, *p* = .005, and confidence, *t*(6355) = 3.52, *p* = .0004. The last letter of the standard identifying code was consistently a predictor of success that was utilized properly for confidence, without bias, and this was the case equivalently across all three groups.

The length of the problem text harmed success when answering on paper (VM1AB and VM1E), in line with numerous previous findings as reviewed above; and confidence reflected this negative effect properly in these groups. When adults worked on screen (VM1F) there was no harmful effect of problem length, with a significant difference from the children's group, VM1AB, *t*(6572) = 2.28, *p* = .02, but not the adult VM1E group, *t*(6915) = 1.45, *p* = .15. However, confidence was still affected by text length, comparably to the other groups. The difference between success and confidence in VM1F reflected oversensitivity of confidence to text length.

Lastly, the inclusion of formulas in the question helped children but not adults. This cue was not utilized by the children (VM1AB), nor by adults working on paper (VM1E), but was negatively utilized by adults working on screen (VM1F). The differences between children and both adult groups were significant for both success, *t*(9925) = 3.72, *p* = .0002, and confidence, *t*(9925) = 2.80, *p* = .005. Notably, the inclusion of a formula generated a bias in all groups, but with different directions. See Figure 10. Such differential effects across populations clearly challenge theoretical explanations. I call for future research to delve into this finding, and consider other biases with such differential effects on different populations.

Figure 10

Experiment 3—Success, confidence, and bias generated by the presence of a formula in the problem across the three groups, VM1AB, VM1E, and VM1F.



In sum, Experiment 3 demonstrates the effectiveness of BEVoCI in shedding new light on previously collected data. Like both previous experiments, these reanalyses reveal impressive cue integration, and several sources for mistakes and confidence biases. Specifically, the detailed examination of confidence exposed it as being oversensitive or undersensitive to some task design features (respectively, text length in VM1F and the presence of a formula among children in VM1AB), while it reliably reflected other aspects of the task (e.g., session number). Moreover, disparities between different populations in their susceptibility to misleading factors is an important aspect of assessment design that is too often overlooked. Calling the attention of learning and exam designers to such biasing factors is a central goal of educational research.

5. General Discussion

The aim of the present study was to introduce the BEVoCI method as an efficient tool for exposing heuristic cues that may mislead people when self-assessing their success in cognitive tasks. I exemplified the methodology using cues for the metacognitive judgment of confidence in problem-solving tasks that downplay the role of vocabulary and semantic skills. Based on insights from previous research, several cue types were examined: cues inherent to the stimuli (e.g., edges and area); stimuli characteristics unrelated to the task goal (e.g., perceived nameability); manipulated cues orthogonal to the goal (e.g., rotation and height); and cues derived from the procedure (serial order). Some have discrete levels (e.g., basic shape area, rotation angle) while others vary over a range of possible values (e.g., perceived nameability, differences in the number of edges) that may be harder for participants to notice.

5.1. Identifying Multiple Sources for Confidence Biases at Once

As explained above, traditionally, researchers have examined single cues in isolation in dedicated studies. This approach can be illustrated by looking at each panel in the figures showing one bias as a main research outcome of this type. In addition, studies which identify double dissociations are rare; and most of those which exist employ two experiments, each dedicated to one direction of the dissociation (e.g., Metcalfe & Finn, 2008; Sidi et al., 2020). Very few studies expose double dissociations using two groups within one sample in the same experiment (e.g., Ackerman & Zalmanov, 2012). Here, each experiment considered several cues as potential biasing factors, some of them with opposite effects on success and confidence, at the same time within one sample.

The present study joins meta-memory studies (Undorf & Bröder, 2021; Undorf et al., 2018) in demonstrating that people can integrate multiple cues at the same time. In particular, in the present study confidence reflected the effects of between two and four cues at the same time. At the same time, the present findings highlight more than before that each considered cue may lead to a unique bias, above and beyond other cues. Moreover, considering several cues at once also supports identifying new cues suspected to affect success and/or confidence on top of already known cues. To illustrate this point, rotation in both the CoP and MTT and height in the CoP affected success in some cases but never misled confidence. These findings alone would not warrant publication, but alongside the myriad biases exposed here, they paint a rich picture of cue integration. Moreover, examining the same cue across different tasks is important for exposing cues that people are particularly attuned to, and those that constitute persistent sources of bias. Both are valuable pieces of information of which educators should be aware.

In all three tasks, the approach of considering multiple cues within one sample exposed double dissociations between effects on success and on confidence. Of course, it is not new that confidence is based on heuristic cues and is reliable only to the extent that the underlying cues predict success (Ackerman, 2019; Koriat, 1997). In line with the classic hard–easy effect (see Suantak et al., 1996), confidence was sometimes, but not always, less sensitive to changes in success than it should be (e.g., formula inclusion in Algebra problems for children but not for adults; see also, underconfidence with practice effect, Koriat et al., 2002). In contrast, confidence was found to be persistently oversensitive to response time—a phenomenon well-established in the metacognitive literature (e.g., Ackerman & Zalmanov, 2012). However, this oversensitivity diminished with the detailed instructions for the CoP task (Exp1c). Notably, the present study highlights an unknown case of oversensitivity of confidence, where basic shape area in the CoP generated a robust bias across five groups of participants that was not overcome with either instructions or background knowledge. This finding deserves further research to help clarify what kinds of cues yield such persistent oversensitivity. Researchers and practitioners may also define thresholds for what level of sensitivity to cues counts or does not count as a bias, depending on the context (e.g., mild bias in utilizing basic shape areas in the CoP; see Table 2 and Figure 5 Panel B).

Notably, though, the analyses performed in the present study investigated cue utilization while assuming that group members integrate cues in a similar manner (see Wiggins & Kolen, 1971 for a similar approach). That is, I did not seek to identify the cues integrated uniquely by each participant. Recent metacognitive studies have shown individual differences in cue integration (e.g., Undorf & Bröder, 2021; Undorf et al., 2018). Differences between theory-guided subgroups may point to additional factors that affect cue integration, such as gender, culture, age, and background knowledge (see Hines et al., 2015, for comparing younger and older adults in meta-memory judgments). The BEVoCI can then be used to examine the differential contribution of each cue and interactions among cues across subgroups. Moreover, clustering participants by their cue integration patterns is yet another direction for future research into heterogeneity (see Cooksey et al., 1986).

Experiment 1 and Experiment 3 included comparisons between populations, incentive structures (bonus schemes), and task-level manipulations (basic vs. detailed instructions, computer vs. paper presentation). Notably, the resolution was persistent within each task across all comparisons. Nevertheless, BEVoCI exposed several factors affecting success and changes in its predictors across conditions. Comparing the effects on confidence revealed that the same cue may act differently in different conditions, both in utilization (whether a cue is utilized or ignored) and in the biases it generates, establishing the malleability of cue integration. Consequently, researchers should be very careful when drawing conclusions regarding biasing factors, as a biasing factor in one condition or for one population may have a different effect in another setting.

Importantly, as demonstrated in Experiment 3, the BEVoCI methodology can be applied to already existing datasets to expose cues and to identify task design and population effects on cue integration. Such discoveries should then be followed up by focused examination, to avoid post-hoc theorizing. By using the BEVoCI, follow-up studies may also expose changes in cue integration with variations of stimuli, task design, and/or population.

The BEVoCI methodology is relevant more generally than demonstrated here. As with the requirements for the dataset chosen for Experiment 3, it can be applied to any task with a clear correct answer for each item (rather than opinion-based or free design tasks) and enough items for each participant to ensure the robustness of the correlation and hierarchical multiple regression analyses at the participant level (ideally several tens of items, but even ten items could be sufficient given enough participants to ensure adequate power). For instance, when teaching a skill during an extended period, serial order effects are the primary measure. However, identifying biasing factors that take effect beyond learning effects is of particular importance. This is demonstrated in the present study by the biases remaining in Experiment 1 even with the solid mathematics background knowledge of the engineering students (Exp1b) and when using detailed instructions (Exp1c), despite pronounced improvement in success (from 53% to 61–71%). In such cases, if there is an improvement in success, biases for confidence and mental effort regulation are rarely considered, although they are prerequisites for long-term achievement (Bjork et al., 2013; de Bruin et al., 2020; Roebers, 2017).

5.2. Fluency and Goal-Oriented Self-Regulation

Response time is often used in metacognitive research as a proxy for fluency. The theorizing behind this is that response time reflects the participant's experience of the item as easy or difficult, which informs how she infers her chance for success (see Baars et al., 2020, for a meta-analysis). Undorf and Bröder (2020) concluded, based on a memorization task, that people take cues into account strategically, rather than bundling them into a single unified feeling of ease or difficulty. In that sense, response time's contribution to cue integration when included in the BEVoCI analyses may reflect what remains after controlling for other considered cues that seem to hint at the difficulty of each task item (Undorf et al., 2022). It then follows that should researchers discover additional cues that contribute to the experience of difficulty, this might weaken the contribution of response time to the explained variance. Moreover, the findings of the present study suggest that cues may affect confidence in varying, and even opposing, directions on the same items, potentially leading to a complex inference of item difficulty that has not previously been recognized.

Although response time reflects the solver's experience of difficulty, it is also affected by strategic motivational variations, such that the higher the motivation, the more time people invest in the task in a goal-driven manner (Ackerman, 2014; Koriat et al., 2006; Undorf & Ackerman, 2017). Moreover, studies have shown that experimental manipulations in the lab, as well as in educational environments, can change whether solvers interpret effort as indicating difficulty, or motivation to succeed (Koriat & Ackerman, 2010; Oyserman et al., 2018; Smith & Oyserman, 2015). Thus, response time as a cue for metacognitive judgments requires a complex inference process. This inference process has been found to develop throughout childhood and adolescence, and into adulthood (Koriat et al., 2014). In light of all these considerations regarding response time, the present study joins Undorf and Bröder (2020) in shaking the ground under the use of response time as a direct cue for metacognitive judgments, as doing so conceals information regarding the underlying bottom-up (cue-based) and top-down (strategic, motivational) processes involved. Instead, it seems necessary to delve into the detailed cues, instructions, and motivations that ultimately underlie the time one invests in each task item (Ackerman, 2014).

5.3. Implications for Instructional Design

Good instructional design can improve outcomes (e.g., Allaire-Duquette et al., 2019; Carney & Levin, 2002; Michalsky, 2021; Sweller et al., 2019) and metacognitive processes (Baars et al., 2013; Carpenter et al., 2019). However, in many cases it is not clear what aspects of the design were effective and which were redundant. The BEVoCI methodology may provide insights into how different elements of instructional design affect learning. In the present study, as demonstrated in all experiments (Figure 5, Figure 8, and Figure 10), the BEVoCI allowed exposing various types of monitoring biases. The detailed instructions in Exp1c attenuated and even eliminated some biases (e.g., difference in edges), but not others. Likewise, the BEVoCI can offer insights into the processes underlying successful and non-successful interventions; and it can be applied to answer questions that go beyond effects on success and general monitoring accuracy, such as whether learners reflect changes in success adequately in their metacognitive judgments. The BEVoCI can thus inform instructional design and interventions by allowing educators to focus improvement efforts on the weak points, where room for improvement is largest and biases are most pronounced, and on those populations which suffer from biases the most.

Finally, assessment theory and practice can benefit from understanding biases in subjective judgments. For instance, Engelhard Jr et al. (2018) considered the effects of various cues on psychometric assessments of essay writing skills. One may also consider how manipulations affect different subjective evaluation contexts, like creativity (Kenett et al., 2021; Sidi et al., 2020), and professional assessments, such as medical skills and decision-making (Beckstead, 2017; Norman & Eva, 2010). Exposing focused biases may be useful in improving the reliability of these important subjective processes.

References

Ackerman, R. (2014). The Diminishing Criterion Model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*(3), 1349-1368.

Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psychological Topics*, *28*(1), 1-20.

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgments during reasoning and memorization. *Thinking & Reasoning*, *23*(4), 376-408.

Ackerman, R., & Leiser, D. (2014). The effect of concrete supplements on metacognitive regulation during learning and open-book test taking. *British Journal of Educational Psychology*, *84*(2), 329-348.

Ackerman, R., Leiser, D., & Shpigelman, M. (2013). Is comprehension of problem solutions resistant to misleading heuristic cues? *Acta Psychologica*, *143*(1), 105-112.

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

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

Ackerman, R., Yom-Tov, E., & Torgovitsky, I. (2020). Using confidence and consensuality to predict time invested in problem solving and in real-life web searching. *Cognition*, *199*, 104248.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187-1192.

Allaire-Duquette, G., Babai, R., & Stavy, R. (2019). Interventions aimed at overcoming intuitive interference: insights from brain-imaging and behavioral studies. *Cognitive processing*, *20*(1), 1-9.

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review*, *33*(2), 693-712.

Ariel, R., Lembeck, N. A., Moffat, S., & Hertzog, C. (2018). Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence*, *70*, 42-51.

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, *33*, 92-107.

Baars, M., Visser, S., Van Gog, T., de Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary educational psychology*, *38*(4), 395-406.

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.

Bates, K. E., & Farran, E. K. (2021). Mental imagery and visual working memory abilities appear to be unrelated in childhood: Evidence for individual differences in strategy use. *Cognitive Development*, *60*, 101120.

Beckstead, J. W. (2017). The bifocal lens model and equation: Examining the linkage between clinical judgments and decisions. *Medical Decision Making*, *37*(1), 35-45.

bin Mohd Noh, M. F., & bin Mohd Matore, M. E. E. (2019). Brunswik's Lens Model: This is how to inspire accurate raters. *Creative Education*, *10*(12), 2859.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417-444.

Bohning, G., & Althouse, J. K. (1997). Using tangrams to teach geometry to young children. *Early childhood education journal*, *24*(4), 239-242.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, *197*, 153-165.

Capraro, M. M., & Joffrion, H. (2006). Algebraic equations: Can middle-school students meaningfully translate from words to mathematical symbols? *Reading Psychology*, *27*(2-3), 147-164.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, *14*(1), 5-26.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51–64.

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.

Castro-Alonso, J. C., Ayres, P., & Paas, F. (2019). VAR: A battery of computer-based instruments to measure visuospatial processing. In *Visuospatial processing for education in health and natural sciences* (pp. 207-229). Springer.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable difficulty effects in the learning of high-element interactivity materials. *Frontiers in Psychology*, 1483.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal*, *23*(1), 41-64.

de Bruin, A. B., 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.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269-273.

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*(25), 23-38.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761-778.

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, *40*(1), 37-55.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271-280.

Engelhard Jr, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, *60*(1), 33-52.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

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.), *The Psychology of Human Thought: An Introduction* (pp. 89-111). Heidelberg University Publishing.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*(443).

Galili, H., Babai, R., & Stavy, R. (2020). Intuitive interference in geometry: An eye-tracking study. *Mind, Brain, and Education*, *14*(2), 155-166.

Golan, D. D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education*, *126*, 346-358.

Halamish, V. (2018). Can very small font size enhance memory? *Memory & cognition*, *46*(6), 979-993.

Händel, M., de Bruin, A. B., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, *15*(1), 51-75.

Hart, Y., Mayo, A. E., Mayo, R., Rozenkrantz, L., Tendler, A., Alon, U., & Noy, L. (2017). Creative foraging: An experimental paradigm for studying exploration and discovery. *PLoS One*, *12*(8), e0182133.

Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*(1), 108-125.

Hines, J. C., Hertzog, C., & Touron, D. R. (2015). Younger and older adults weigh multiple cues in a similar manner to generate judgments of learning. *Aging, Neuropsychology, and Cognition*, *22*(6), 693-711.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological bulletin*, *134*(3), 404.

Kaufmann, E. (2022). Lens model studies: Revealing teachers' judgements for teacher education. *Journal of Education for Teaching*, 1-16.

Kenett, Y. N., Rosen, D. S., Tamez, E. R., & Thompson-Schill, S. L. (2021). Noninvasive brain stimulation to lateral prefrontal cortex alters the novelty of creative idea generation. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(2), 311-326.

Kleitman, S., & Moscrop, T. (2010). Self-confidence and academic achievements in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender. In A. Efklides & P. Misailidi (Eds.), *Trends and Prospects in Metacognition Research. Part 2* (pp. 293-326). Springer.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349-370.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong? *Journal of Experimental Psychology: General*, *47*(5), 613-631.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, *19*(1), 251-264.

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

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development*, *24*(2), 169-182.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 34-53.

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), 36-68.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology-General*, *131*(2), 147-162.

Kornell, N., & Hausman, H. (2017). Performance bias: Why judgments of learning are not affected by learning. *Memory & cognition*, *45*(8), 1270-1280.

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, *35*, 455-463.

Leahy, W., & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. *Instructional science*, *44*(1), 107-123.

Lee, J., Lee, J. O., & Collins, D. (2009). Enhancing children's spatial sense using tangrams. *Childhood Education*, *86*(2), 92-94.

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & cognition*, *43*(4), 681-693.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174-179.

Michalsky, T. (2021). When to scaffold motivational self-regulation strategies for high school students' science text comprehension. *Frontiers in Psychology*, *12*, 658027.

Morsanyi, K., Prado, J., & Richland, L. E. (2018). The role of reasoning in mathematical thinking. *Thinking & Reasoning*, *24*(2), 129-137.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*(20), 378-384.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical education*, *44*(1), 94-100.

Oyserman, D., Elmore, K., Novin, S., Fisher, O., & Smith, G. C. (2018). Guiding people to interpret their experienced difficulty as importance highlights their academic possibilities and improves their academic performance. *Frontiers in Psychology*, *9*, 781.

Parsons, L. M. (1995). Inability to reason about an object's orientation using an axis and angle of rotation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1259.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2019). R Core Team. 2019. nlme: linear and nonlinear mixed effects models. R package version 3.1-141. *Available at h Ttp://CRAN. R-Project. Org/Package= Nlme*.

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J., & Van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning*, *14*(1), 21-42.

Rau, M. A., & Matthews, P. G. (2017). How to make ‘more’ better? Principles for effective use of multiple representations to enhance students’ learning about fractions. *ZDM*, *49*(4), 531-544.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615-625.

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental review*, *45*, 31-51.

Roll, I., Baker, R. S. d., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, *23*(4), 537-560.

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.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, *18*(7), 698-711.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Sidi, Y., Ophir, Y., & Ackerman, R. (2016). Generalizing screen inferiority-does the medium, screen versus paper, affect performance even with brief tasks? *Metacognition and Learning*, *11*(1), 15-33.

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.

Sidi, Y., Torgovitsky, I., Soibelman, D., Miron-Spektor, E., & Ackerman, R. (2020). You may be more original than you think: Predictable biases in self-assessment of originality. *Acta Psychologica*, *203*, 103002.

Smith, G. C., & Oyserman, D. (2015). Just not worth my time? Experienced difficulty and time investment. *Social cognition*, *33*(2), 85-103.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). Toward a rationality quotient (RQ): The comprehensive assessment of rational thinking (CART). In *The Thinking Mind* (pp. 216-236). Psychology Press.

Stavy, R., & Babai, R. (2008). Complexity of shapes and quantitative reasoning in geometry. *Mind, Brain, and Education*, *2*(4), 170-176.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*(2), 201-221.

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*, 973-978.

Sweller, J., Merriënboer, J. J. G. v., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261-292.

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger III, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition*, *7*(1), 106-115.

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.

Thiede, K. W., Wright, K. L., Hagenah, S., Wenner, J., Abbott, J., & Arechiga, A. (2022). Drawing to improve metacomprehension accuracy. *Learning and Instruction*, *77*, 101541.

Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*, 237-251.

Tindall-Ford, S., Agostinho, S., & Sweller, J. (2020). *Advances in cognitive load theory*. London: Routledge.

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.

Träff, U., Olsson, L., Skagerlund, K., Skagenholt, M., & Östergren, R. (2019). Logical Reasoning, Spatial Processing, and Verbal Working Memory: Longitudinal Predictors of Physics Achievement at Age 12–13 Years. *Frontiers in Psychology*, *10*, 1929.

Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review*, *24*(6), 2003-2011.

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629-642.

Undorf, M., & Bröder, A. (2021). Metamemory for pictures of naturalistic scenes: Assessment of accuracy and cue utilization. *Memory & cognition*, 1-18.

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & cognition*, *43*(4), 647-658.

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me"–Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*, 105011.

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & cognition*, *46*(4), 507-519.

Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 97-109.

Vuorre, M., & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*, *17*(2), 269-291.

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, *107*(4), 1051.

Wiggins, N., & Kolen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918-933.

Zak, Y., Tapiro, H., Alicia, T. J., Parmet, Y., Rottem Hovev, M., Taylor, G. S., & Oron-Gilad, T. (2021). Rapid Interpretation of Temporal–Spatial Unmanned Aerial Vehicle (UAV) Operational Data–RITSUD: Aiding UAV Operators With Visualizations of Patterns-of-Life Activities. *Journal of Cognitive Engineering and Decision Making*, 15553434211023605.