Judging the accuracy of eyewitness testimonies using retrieval effort cues

Philip U. Gustafsson | Torun Lindholm | Fredrik U. Jönsson

Department of Psychology, Stockholm University, Stockholm, Sweden

Correspondence
Philip U. Gustafsson, Department of Psychology, Stockholm University, Stockholm 106 91, Sweden.
Email: philip.gustafsson@psychology.su.se

Funding information
Elisabeth and Herman Rhodin Memorial Foundation; The Lars Hierta Memorial Foundation

Summary
Recent research has shown that incorrect statements in eyewitness testimonies contain more cues to effortful memory retrieval than correct statements. In two experiments, we attempted to improve judgments of testimony accuracy by informing participants about these effort cues. Participants read eyewitness testimony transcripts and judged statement accuracy. Performance was above chance in both experiments, but there was only a significant effect of the effort-cue instruction in Experiment 2. In Experiment 1, we also compared judgment accuracy between police detectives, police students and laypersons, and found no significant difference, in contrast to previous studies. Moreover, the current study corroborates previous findings that (a) judging testimony accuracy is a difficult task and (b) people spontaneously rely on effort cues to some extent when judging accuracy. However, a complete reliance on effort cues showed substantially better performance than relying on one’s own judgments skills at best, and offered equal performance at worst.

KEYWORDS
eyewitness accuracy, eyewitness memory, eyewitness testimony, judgment accuracy, retrieval effort cues

1 | INTRODUCTION

Judging the accuracy of another person’s memory is not always a trivial matter. In courts all over the world, judges and jurors must decide whether eyewitness testimonies are correct, sometimes with the fate of people’s lives at stake. Indeed, eyewitness testimonies are often critical in crime investigations and ensuing trials (Wells et al., 2006). However, eyewitness evidence can be unreliable, as some witnesses lie (see DePaulo et al., 2003; Sporer & Schwandt, 2006; Vrij et al., 2017)—and perhaps more commonly—provide honest but erroneous accounts. The fallibility of eyewitness evidence becomes strikingly clear by estimations that more than 70% of overturned wrongful convictions have involved eyewitness misidentifications (Innocence Project, 2021, see also Garrett, 2011). However, the negative consequences that may follow from erroneous recall can be avoided if the error is detected. It is therefore crucial that fact finders obtain the skill to separate correct from incorrect memories.

Most research on eyewitness accuracy has focused on delineating accuracy in eyewitness identifications (i.e., picking out a suspect in a lineup; see Wells et al., 2006 for an overview). However, mistaken identification is not the only aspect of eyewitness memory of importance for the legal system. At least as important for the outcome of a legal trial is witnesses’ verbal testimony. Existing research on evaluations of verbal eyewitness testimonies overall demonstrate accuracy rates only slightly above chance level (Ball & O’Callaghan, 2001; Clark-Foos et al., 2015; Johnson & Suengas, 1989; Lindholm, 2005, 2008a, 2008b; Schooler et al., 1986).

Recently, certain cues in statements from verbal eyewitness testimonies have been found to relate to memory accuracy (Gustafsson et al., 2019; Lindholm et al., 2018). These markers are based on a
witness' verbal and paraverbal expressions of effort when attempting to remember an event. In these studies, witnesses expressed markers of effort, like pausing, hedging (e.g., “maybe,” “possibly”), and using interjections such as “eh” “uh”; more often when the details they recalled were incorrect rather than correct. The current study will investigate whether it is possible to improve judgment accuracy of verbal eyewitness reports, by informing fact finders about the effort cues—accuracy relation.

1.1 | Memory accuracy

A fundamental question for addressing the ability to judge the accuracy of memories is whether there are any qualitative differences between correct and incorrect memories. One line of research, reality monitoring (Johnson & Raye, 1981; later “source monitoring,” Johnson et al., 1993) have proposed that differences should ensue between real (“correct”) and imagined (“incorrect”) memories as they are generated from different sources. Specifically, real memories are proposed to be mainly generated from external information—such as sensory, spatial and temporal information—and be picked up by perceptual input. Contrastingly, imagined memories are proposed to mainly be generated from internal information, such as cognitive processes. Following these sources, real memories should, according to reality monitoring theory, contain more sensory, spatial and temporal information, whereas imagined memories should contain more information about cognitive operations, such as reflections on the experience of retrieving a memory. Research supports this idea (Hashtroudi et al., 1990; Kensinger & Schacter, 2006; Schooler et al., 1986; Sporer & Sharman, 2006; Strömwall & Granhag, 2005).

Cue utilization theory is another prominent theory that also suggests qualitative differences between correct and incorrect memories (see Koriat, 1997, 2006). The theory proposes that we rely on cues relating to memory processes when we judge our memories. For example, easily and quickly recalled memories tend to be judged as more accurate than those that are difficult to recall (Kelley & Lindsay, 1993; Robinson et al., 1997). Moreover, research shows that correct memories are retrieved quicker than incorrect memories (Ackerman & Koriat, 2011; Brewer et al., 2006; Brewer & Weber, 2008; Koriat & Ackerman, 2010; Smith & Clark, 1993; Weidemann & Kahana, 2016). Thus, the memory retrieval process, such as the ease with which a memory is retrieved, can give an indication of a memory’s accuracy. Further evidence that retrieval ease, or phrased differently, retrieval effort, is related to memory accuracy, has been found in two recent studies in an eyewitness context. Lindholm et al. (2018) and Gustafsson et al. (2019) showed participants a mock crime film and then interviewed them as eyewitnesses. The statements given by the witnesses in these interviews were then coded for accuracy and markers of effort (effort cues). The results showed that incorrect statements contained more effort cues, such as more delays in a response, more fillers (e.g., “eh,” “let me see”) and more hedges (e.g., “I’m not sure,” “maybe”), compared to correct responses.

Thus, research supports the idea that correct and incorrect memories differ. However, do people use this information when judging other's memory accuracy, and to what extent are people's methods successful?

1.2 | Judging the accuracy of others' memories

Reality monitoring theory has been used in investigations of judgments of others’ memories. For example, Johnson et al. (1998) manipulated perceptual and emotional details in a story of an autobiographical recollection and asked fact finders to judge whether the story was experienced first-hand (“real”) or retold from someone who only heard about it (“imagined”). They found that fact finders were more likely to judge a memory as real when the memory contained more perceptual and emotional details (see also Keogh & Markham, 1998; Sporer & Sharman, 2006). Clark-Foos et al. (2015) carried out a similar study, investigating whether the accuracy of interpersonal judgments could improve with training and feedback. They found the accuracy of fact finders’ judgments to be above chance level, and that performance improved with feedback and training. However, the real memories actually contained fewer spatial and sensory details than the imagined memories, counter to reality monitoring theory. In another similar study by Johnson and Suengas (1989), participants acting as witnesses were asked to either recollect or make up (i.e., imagine) a memory, and focus on either perceptual or apperceptive (e.g., emotional, cognitive) details of that memory. Fact finders then had to judge which of the memories were real and which were imagined. Results showed the fact finders generally judged witness reports high in perceptual information as likely to be real, in line with the reality monitoring theory. However, as both the perceptual-focusing and the apperceptive-focusing witnesses had talked about their memories in these respective terms—regardless of being real or imagined—the resulting accuracy of the fact finders reality judgments were at chance level. That is, fact finders relied on perceptual detail for reality judgments, but it was not an effective method (although this was largely due to the experimental setup). Taken together, these studies demonstrate some difficulties in using reality monitoring to delineate eyewitness accuracy.

Other studies looking at judgments of others’ memories have had a cue-utilization perspective. Jameson et al. (1993) investigated people’s judgments of someone else’s knowledge, so-called “feeling of another person’s knowing” (FOAK). One group of participants was tasked with answering general knowledge questions and made predictions about the likelihood of recognizing the answer if given the opportunity, followed by a multiple-choice test. Another group of fact finders observed the earlier participants’ failed recall attempts, and then judged the likelihood that the participants would recognize the answer if it was shown to them. The fact finders accurately gave higher FOAK-judgments for trials where the participants also made higher predictions of later recognition. Importantly, fact finders’ FOAK-judgments positively correlated with response latency, in that the longer participants searched for an answer (albeit unsuccessfully), the higher the fact finders FOAK-judgments. Brennan and Williams (1995) later replicated the latter finding, but they also found the reverse relationship participants when participants did provide an answer, that is, their FOAK-judgments were higher the quicker the answer. This indicates that fact finders’ judgments were cue-based.
Although the studies of judgments of others’ semantic memory above have examined judgments of individual statements, such as the answer to specific questions, studies of judgments of others’ episodic memory have largely focused on the credibility of eyewitness testimonies as a whole (Bell & Loftus, 1989; Borckardt et al., 2003; Cutler et al., 1989; Lindholm, 2008a). Research on judgments of others’ episodic memory on a more detailed and specific, statement-level is more scarce. It is important to be able to accurately judge individual statements in eyewitness settings, such as when compiling a description of an offender, as an eyewitness may be correct about certain details (e.g., clothing) but incorrect about others (e.g., facial features).

In one of the few studies that investigated judgments of individual episodic memories, Ball and O’Callaghan (2001) first observed children at a dentist and then interviewed them about the visit. Fact finders then judged the accuracy of the children’s answers, and although they could distinguish correct answers from incorrect answers, the accuracy of their judgments was only slightly above chance level. Lindholm (2008b) carried out a similar study in a legal setting. Eyewitnesses first watched a video of a kidnapping and were interviewed about it. Fact finders then judged the accuracy of excerpted statements from the interviews, either by reading or watching them. Three different groups with varying level of law enforcement-experience were tested: police detectives, judges, and laypersons. Lindholm (2008b) found that only the police detectives could consistently detect accuracy in the testimonies, with judges and laypersons performing near, or at, chance level.

The at-most moderate success of fact finders’ judgments of others’ memories on a statement level suggest that there is room for improvement. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve their memory retrieval effort cues (as found in the studies by Lindholm 2008b) aim to improve the accuracy of fact finders’ judgments of others’ memories appear to differ in certain aspects, training fact finders to improve. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve. Together with evidence that correct and incorrect memories appear to differ in certain aspects, training fact finders to improve.
in the mock crime films. Next, two coders blind to the purpose of the study coded the accuracy of the statements (interrater reliability \( r = .75 \)). Then, two new coders selected statements from answers to the recalled questions that were either correct or incorrect (interrater reliability \( r = .95 \)). Statements that were only partly correct were excluded. Finally, effort cues in statements were picked out. Three markers of effortful retrieval were coded (again with a new set of blind coders): Hedges—uncertainty and commitment avoidance such as “maybe,” “possibly” (interrater reliability Cohen’s \( \kappa = .87 \), exact overlap = 62%); Non-word fillers—expressions such as “uh,” “hm” (interrater reliability Cohen’s \( \kappa = .97 \), exact overlap = 91%); and Word fillers—self-talk such as “let me see,” and “meaningless” words such as “well” (interrater reliability Cohen’s \( \kappa = .83 \), exact overlap = 65%). Moreover, two effort cues—response latency and delays—were measured by calculating the elapsed time of silence between utterances. Response latency was operationalized as the elapsed silence before the start of a statement. Delays were operationalized as elapsed silence of at least 2 s before, or during, a statement (in contrast to shorter silences that always arise between a question and an answer). For a full description of the procedure, see Gustafsson et al. (2019). In Gustafsson et al. (2019), Hedges and Delays proved the strongest (and unique) predictors of accuracy, and were therefore selected as effort cues to be used in this study.

The final transcribed documents contained the full verbatim interview, with the cataloged statements marked in boldfaced characters. Marked pauses in a statement were noted with numbers in parentheses that indicated a pause of that many seconds. Each statement was also preceded with a number that acted as a unique identifier for that statement.

In the current experiment, participants were randomly assigned to either the effort-cue instruction condition or control condition and were handed the transcribed document together with instructions and an answer sheet. The instructions briefly informed the participant of the crime and interrogation of the eyewitness. The participant was then informed how to read the testimony: boldfaced statements were to be judged for accuracy, numbers within parentheses indicated a pause of that many seconds, and the number preceding each statement was its unique identifier. The identifier corresponded to an identical number in the answering sheet, where the participants were instructed to circle the word “correct” or “incorrect,” based on their judgment of the respective statement. Participants in the effort-cue instruction condition were informed that incorrect memories are more often produced with delays longer than 2 s, and more often contain hedges, compared to correct memories (as evidenced from the results in Gustafsson et al., 2019). They were then instructed to take this information into account when judging the statements in the testimony. Participants in the control condition did not receive this information and were instead instructed to judge the statements based on their own knowledge and beliefs about what differentiates correct and incorrect memories. To give participants a sense of the eyewitness’ general style of expressing him/herself, all participants were asked to read through the entire testimony before judging the accuracy of the individual statements.

All 22 testimonies from Gustafsson et al. (2019) were used. Each participant was randomly assigned one testimony, which was balanced across manipulations. The testimonies contained an average of 47.41 statements (\( SD = 10.62 \)), with a mean accuracy of 75.68% (\( SD = 6.19\% \)).

### 2.2 Results

#### 2.2.1 Accuracy judgments

Signal detection theory (Tanner Jr & Swets, 1954) was used to calculate judgment accuracy. A correction factor of 0.01 was added to hit-and false alarm rates of zero, and –0.01 for rates of 1, in order to avoid infinite values (Macmillan & Creelman, 2004). Mean \( d' \) values, as well as hit- and false alarm rates are presented in Table 1.

We first used one sample \( t \)-tests to examine if judgment accuracy (\( d' \)) in the different groups differed from chance level (\( d' = 0 \)). All groups performed significantly above chance: effort-cue instruction condition \( (M = 0.54, SD = 0.58, t(65) = 7.51, p < .001, d = 0.93) \); control condition \( (M = 0.45, SD = 0.55, t(63) = 6.57, p < .001, d = 0.82) \); police detectives \( (M = 0.56, SD = 0.55, t(44) = 6.90, p < .001, d = 1.02) \); police students \( (M = 0.58, SD = 0.51, t(40) = 7.26, p < .001, d = 1.14) \); and laypersons \( (M = 0.36, SD = 0.62, t(43) = 3.78, p < .001, d = 0.58) \).

We next examined the relative performance of each group. A 2 (Instruction: effort cues/control) by 3 (Group: police detectives/police students/laypersons) between-participants ANOVA with \( d' \) as the outcome variable showed no statistically significant effect of instruction, \( F(1, 130) = 0.78, p = .378, \eta^2 = .006 \) (see Figure 1), nor of group \( F(1, 130) = 2.88, p = .092, \eta^2 = .022 \) on judgment accuracy. The

| TABLE 1 | Means (and standard deviations) for hit rate, false alarm rate, discrimination accuracy (\( d' \)) and response bias (\( c \)) in the different groups |
|-------------------|-------------------|-------------------|-------------------|
| Effort cue instruction condition | Control condition |
| | Police detectives | Police students | Laypersons | Police detectives | Police students | Laypersons |
| **Hit rate** | | | | | | |
| Police detectives | 0.70 (0.15) | 0.70 (0.12) | 0.72 (0.12) | 0.69 (0.09) | 0.71 (0.11) | 0.70 (0.14) |
| Police students | 0.52 (0.23) | 0.46 (0.19) | 0.56 (0.18) | 0.47 (0.20) | 0.53 (0.19) | 0.60 (0.22) |
| Laypersons | 0.53 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| **False alarm rate** | | | | | | |
| Police detectives | 0.32 (0.55) | 0.21 (0.36) | 0.42 (0.41) | 0.22 (0.35) | 0.34 (0.38) | 0.48 (0.61) |
| Police students | 0.55 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| Laypersons | 0.53 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| **\( d' \)** | | | | | | |
| Police detectives | 0.52 (0.23) | 0.46 (0.19) | 0.56 (0.18) | 0.47 (0.20) | 0.53 (0.19) | 0.60 (0.22) |
| Police students | 0.53 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| Laypersons | 0.53 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| **\( c \)** | | | | | | |
| Police detectives | -0.32 (0.55) | -0.21 (0.36) | -0.42 (0.41) | -0.22 (0.35) | -0.34 (0.38) | -0.48 (0.61) |
| Police students | 0.55 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
| Laypersons | 0.53 (0.56) | 0.67 (0.55) | 0.44 (0.64) | 0.59 (0.54) | 0.50 (0.47) | 0.26 (0.61) |
interaction between instruction and group was also statistically non-significant, F(1, 126) = 0.95, p = .332, η² = .007.

**Effort observer.** To examine if participants in the effort-cue instruction condition had followed the instructions, we compared their results with a computed “effort observer” that judged the statements from all testimonies completely based on the effort instruction. That is, each statement that contained no hedges or pauses was coded as correct, and all other statements were coded as incorrect. A \( d' \) score was then calculated for each testimony. The mean judgment accuracy \( d' \) of the “effort observer” (\( M = 0.79, SD = 0.65 \)) was higher than for the effort-cue instruction group (\( M = 0.54, SD = 0.58; d = 0.77 \)), see Figure 1.

### 2.2.2 | Response bias

Next, we explored participants’ response bias \( c \), that is, the general tendency to judge statements as correct rather than incorrect.\(^3\) Mean values for \( c \) are presented in Table 1.

We first examined if each group were biased by comparing their scores to no bias \( c = 0 \) with one sample \( t \)-tests, and all groups had a liberal bias: **effort-cue instruction condition** \( M = -0.33, SD = 0.46 \), \( t(65) = 5.79, p < .001, d = 0.72 \); **control condition** \( M = -0.35, SD = 0.47 \), \( t(63) = 5.90, p < .001, d = 0.74 \); **police detectives** \( M = -0.27, SD = 0.47 \), \( t(44) = 3.90, p < .001, d = 0.57 \); **police students** \( M = -0.28, SD = 0.38 \), \( t(40) = 4.83, p < .001, d = 0.74 \); and **laypersons** \( M = -0.45, SD = 0.51 \), \( t(43) = 5.81, p < .001, d = 0.88 \).

We next examined the relative bias of each group. The instruction by group ANOVA with \( c \) as the outcome variable showed no statistically significant effect of instruction, \( F(1, 130) = 0.73, p = .787, η² < .001 \), nor of group \( F(1, 130) = 3.29, p = .072, η² = .025 \) on response bias.

### 2.2.3 | Judgment accuracy of individual witnesses

The witnesses whose statements were included in this study varied in the extent to which they used hedges or how quickly they responded. To examine the extent to which information about effort cues could be used to reliably predict accuracy, we explored the correlation between amount of effort cues in each individual witness testimony and judgment accuracy. Specifically, we constructed a ratio of how many more effort cues were present in incorrect responses relative to correct responses (i.e., \( \text{effort cues}_{\text{incorrect}} / \text{effort cues}_{\text{correct}} \)). Mean effort ratio for testimonies was 2.19 (SD = 1.15). That is, slightly over twice as many effort cues occurred in incorrect statements compared to correct statements. We next carried out Pearson’s correlations between judgment accuracy \( d' \) and effort ratio for each instruction group (effort-cue instruction/control), as well as for the “effort observer.” Results showed a statistically significant correlation between judgment accuracy and testimony effort ratio for the effort-cue instruction group, \( r(64) = .41, p < .001 \) and also for the control group, \( r(62) = .51, p < .001 \), as well as for the effort observer, \( r(20) = .63, p = .002 \). Thus, as the relative amount of effort cues in incorrect testimonies increased, so did the accuracy of participants’ judgments in both groups, as well as for the effort observer.

### 2.3 | Discussion

Overall, all groups performed slightly above chance in judging accuracy, on levels largely comparable to previous studies (Ball & O’Callaghan, 2001; Clark-Foos et al., 2015; Lindholm, 2008b). Contrary to expectations, participants in the effort cue-condition were not significantly better at judging accuracy compared to those in the control condition (see Figure 1). Such a result is surprising, given the previously established relationship between memory retrieval effort and accuracy (Gustafsson et al., 2019; Lindholm et al., 2018). A straightforward explanation for this result could be a lack of power, as the a priori power analysis suggested a slightly underpowered study at 77.50% power. However, an alternative explanation could be that the effort cue instruction group only partially relied on the instructions, and combined them with other cues. Support for this is found by comparing the effort cue groups result with a computation for what the results would have been, should the participants have fully relied on information about the effort cues (i.e., the “effort observer,” see Figure 1). Although no significance test was carried out—due to each testimony only being judged once by the “effort observer”—a comparison of means shows that the effort observer had a higher score than the effort instruction group, in terms of a large effect size \( d = 0.77 \); see Cohen, 1988).

Another unexpected result was that police detectives were not significantly more accurate in their judgments than the presumably less experienced police students and laypersons, contrary to our hypothesis and previous studies (Lindholm, 2008b; Lindholm et al., 1997). A limitation in this experiment is that we did not obtain information regarding the amount of experience the police detectives
had working in the law-enforcement business. Thus, our sample may have included several inexperienced detectives, which could explain the discrepancy between our result and previous studies. Another explanation is a lack of power, as the a priori calculation only suggested 67.56% power. Furthermore, we hypothesized that the police students and laypersons would improve in the effort-cue instruction condition, but that police detectives would not. In numbers, results showed that judgment accuracy was indeed higher for both the police students and the laypersons in the effort instruction group compared to the control group, whereas the reverse was true for the police detectives (see Table 1). However, these differences were not statistically significant and effects small overall. Again, a somewhat low power (67.53%) could account for these results, and a larger sample may have yielded a greater effect.

All groups had a slight liberal bias—that is, a greater tendency to judge a statement as correct rather than incorrect—but there were no statistically significant differences between the groups. These results replicate findings in previous studies (Ball & O’Callaghan, 2001; Lindholm, 2008b), albeit less pronounced in this experiment. Of note is that this bias is approximate to the mean accuracy rate of the testimonies (75.68%), which possibly indicates that participants had a good “feel” of the witnesses’ general memory performance, even though judgment accuracy of the individual statements were modest.

We also explored judgment accuracy for the different testimonies, in which we found a positive trend. Testimonies with greater effort ratios, that is, more effort cues in incorrect responses relative to correct responses, were more accurately judged. This was the case both for the effort-cue instruction group and the control group, as well as the “effort observer.” One interpretation of these findings is that people might spontaneously rely on effort cues when judging accuracy, which is supported by results from Jameson et al. (1993) and Brennan and Williams (1995). Somewhat perplexingly, the effect was even greater for the control group compared to the effort-cue instruction group. However, these results are exploratory and therefore need further validation before drawing any strong conclusions.

Taken together, the current experiment found no advantage in getting instructed about effort cues to judge accuracy compared to a reliance on one’s own methods. However, the results from the “effort observer” indicate that participants might not use this information optimally.

3 | EXPERIMENT 2

In Experiment 1, we found that participants in the effort-cue instruction condition did not perform as well they potentially could have, if they had instead made their judgments completely based on the effort cues. In this follow-up experiment, we therefore wanted to clarify the instructions, to ensure that participants used our suggested cues to judge eyewitness testimonies. Furthermore, we wanted to increase the reliability of judgments of each witness and therefore decided to use three testimonies as stimulus material, instead of the full 22 used in the first experiment. We again hypothesized that the group instructed to use effort cues when judging testimonies would be more accurate than the control group. Furthermore, based on results in Experiment 1, we hypothesized that the testimony with the largest effort ratio would be judged most accurately, followed by the testimony with the medium effort ratio, and lastly the testimony with the smallest effort ratio. We also hypothesized an interaction, namely that the difference between the effort cue-instruction group and the control group would be larger the higher the effort ratio, in favor of the effort-cue instruction group. This experiment has been preregistered (https://osf.io/4c58n).

3.1 | Method

3.1.1 | Participants and design

A 2 (instruction: effort cues/control) × 3 (testimony effort ratio: small/medium/large) between-participants design was used. A power analysis suggested a total of 158 participants to obtain 80% power. We recruited a total 208 participants, but 46 of these failed to complete a control question at the end of the survey and were subsequently excluded. This left us with a sample of 160 participants, of which 82 made up the effort-cue instruction group (30% men; mean age = 30.94, SD = 10.40), while 78 participants made up the control group (26% men; mean age = 29.04, SD = 8.91). Recruitment sites were identical to the sites for laypersons in Experiment 1. People that had previously participated in studies as eyewitnesses or had previously judged eyewitness testimonies were denied participation. The experiment was conducted in accordance with the ethical principles outlined in ALLEA (2017) and did not include factors that require ethical vetting according to Swedish legislation on research ethics SFS (2003).

3.1.2 | Materials and procedure

Three testimonies from Experiment 1 was used in this experiment. These testimonies were chosen as they encompassed the effort ratio range of the 22 testimonies used in Experiment 1 (see Results in Experiment 1). The small effort-ratio witness testimony represented the bottom end of the effort ratio spectrum (34 statements; accuracy rate = 85.29%), with an effort ratio of 0.46. This means that incorrect statements contained slightly less than half as many effort cues compared to correct statements. The medium effort-ratio witness testimony was the testimony closest to the mean effort ratio (39 statements; accuracy rate = 71.79%), with an effort ratio of 2.27. That is, incorrect statements contained slightly over twice as many effort cues compared to correct statements. The large effort-ratio witness testimony represented the top end of the effort ratio spectrum (39 statements; accuracy rate = 82.05%), with an effort ratio of 4.46, meaning that incorrect statements contained over four times as many effort cues as the correct statements.
The experimental procedure was similar to Experiment 1, but with the following modifications: First, the instructions given to the participants in the effort-cue instruction group were made more direct. We instructed them to make their accuracy judgments completely based on the prevalence or absence of delays and hedges. That is, if any statement contained at least one pause, one hedge, or both, it were to be judged as incorrect. All other statements were to be judged as correct. Second, participants received this instruction before each statement was shown (compared to only once in the introductory information in Experiment 1). That is, participants in the effort-cue instruction group received information to judge accuracy based on delays and hedges, whereas the control group received information to judge accuracy based on their own knowledge and beliefs. Third, the experiment was conducted online, instead of the pen-and-paper approach used in Experiment 1.

3.2 | Results

3.2.1 | Accuracy judgments

Signal detection theory was again used to calculate hit rates, false alarm rates, accuracy (d') and bias (c), with corrections of 0.01 and –0.01 added to hit- and false alarm rates of 0 and 1, respectively (Macmillan & Creelman, 2004). Mean values for d' and c, as well as hit- and false alarm rates, are presented in Table 2.

We first examined if each groups’ judgment accuracy (d') differed from chance level (d' = 0) with one sample t-tests. Results showed that the effort instruction group scored significantly above chance (M = 1.09, SD = 1.01), t(81) = 9.73, p < .001, d = 1.08, whereas the control group did not (M = 0.18, SD = 0.92), t(77) = 1.70, p = .094, d = 0.20. Judgment accuracy for the different testimonies showed accuracy rates significantly above chance for the large effort-ratio testimony (M = 1.01, SD = 1.04), t(49) = 6.88, p < .001, d = 0.97, and the medium effort-ratio testimony (M = 1.11, SD = 0.74), t (56) = 11.43, p < .001, d = 1.50, but not for the small effort-ratio testimony (M = –0.21, SD = 0.87), t(52) = –1.76, p = .084, d = 0.23.

We next examined the relative performance of each group. A 2 (instruction: effort cues/control) by 3 (testimony effort ratio: small/medium/large) between-participants ANOVA with d' as the outcome variable showed a statistically significant effect of instruction, F(1, 156) = 50.64, p < .001, η² = .180, in which participants in the effort-cue instruction group made more accurate judgments compared to the control group (see Figure 2). There was also a statistically significant effect of testimony effort ratio F(1, 156) = 59.38, p < .001, η² = .215. The medium effort-ratio testimony was judged most accurately followed by the large effort-ratio testimony, while the small effort-ratio testimony was judged least accurately. The interaction between instruction and testimony effort ratio was also statistically significant, F(1, 156) = 10.34, p = .002, η² = .037.

Planned comparisons (see preregistration) with Bonferroni corrections revealed that the effort-cue instruction group made significantly more accurate judgments compared to the control group when judging the large effort-ratio testimony (Meff = 1.36, p < .001, d = 1.74), and when judging the medium effort-ratio testimony (Meff = 1.04, p < .001, d = 2.02), but not when judging the small effort-ratio testimony (Meff = .33, p = .172, d = .38), see Table 2. Furthermore, participants judging the small effort-ratio testimony made significantly less accurate judgments compared to both participants judging the medium effort-ratio testimony (p < .001, d = 1.27). There was no statistically significant difference between the participants judging the medium and large effort-ratio testimonies (p = 1.0, d = 0.11).

3.2.2 | Response bias

Next, we explored response bias. We first examined if each group were biased by comparing their scores to no bias (c = 0) with one sample t-tests. Results showed that the effort-cue instruction group had a conservative bias (M = 0.54, SD = 0.75), t(81) = 6.54, p < .001, d = 0.72, whereas the control group had a liberal bias (M = –0.51, SD = 0.72), t(77) = –6.22, p < .001, d = 0.71. The small effort-ratio testimony (M = –0.53, SD = 0.65) was judged with a liberal bias, t (52) = –6.00, p < .001, d = 0.82 whereas the medium effort-ratio testimony (M = 0.45, SD = 0.81) was judged with a conservative bias, t (56) = 4.17, p < .001, d = 0.56. The large effort-ratio testimony

### Table 2

| Effort cue instruction condition | Control condition |
|---------------------------------|-------------------|
|                                 | Small effort ratio testimony | Mid effort ratio testimony | Large effort ratio testimony | Small effort ratio testimony | Mid effort ratio testimony | Large effort ratio testimony |
| Hit rate                        | 0.59 (0.10)        | 0.41 (0.11)        | 0.50 (0.12)                  | 0.72 (0.12)                  | 0.66 (0.15)                  | 0.72 (0.19)                  |
| False alarm rate                | 0.59 (0.18)        | 0.08 (0.13)        | 0.12 (0.19)                  | 0.74 (0.28)                  | 0.47 (0.23)                  | 0.61 (0.26)                  |
| d'                              | –0.05 (0.52)       | 1.63 (0.53)        | 1.67 (0.80)                  | –0.38 (1.12)                 | 0.58 (0.51)                  | 0.30 (0.77)                  |
| c                               | –0.26 (0.39)       | 1.04 (0.50)        | 0.83 (0.58)                  | –0.82 (0.74)                 | –0.17 (0.56)                 | –0.57 (0.73)                 |
Box plots of judgment accuracy for the control group and effort-cue instruction group in Experiment 2. Filled dots show group mean. Bold lines show group median. The dotted horizontal line indicates chance performance.

We then examined the relative bias of each group. An exploratory analysis (instruction: effort cues/control) by 3 (testimony effort ratio: small/medium/large) between-participants ANOVA with c as the outcome variable showed a statistically significant effect of instruction, $F(1, 156) = 99.56, p < .001, \eta^2 = .337$, in which participants in the effort-cue instruction group were more conservative in their judgments compared to the control group. There was also a statistically significant effect of testimony effort ratio, $F(1, 156) = 27.53, p < .001, \eta^2 = .142$. The medium effort-ratio testimony was judged most conservatively, followed by the large effort-ratio testimony, while the small effort-ratio testimony was judged most liberally. The interaction between instruction and testimony effort ratio was also statistically significant, $F(1, 156) = 9.90, p = .002, \eta^2 = .034$.

Exploratory post-hoc tests with Bonferroni corrections revealed that the effort-cue instruction group was more conservative than the control group when judging all testimonies (small effort-ratio testimony $M_{\text{eff}} = -0.56, p < .001, d = 0.95$; medium effort-ratio testimony $M_{\text{eff}} = -1.21, p < .001, d = 2.28$; large effort-ratio testimony $M_{\text{eff}} = -1.40, p < .001, d = 2.12$), see Table 2. Furthermore, the small effort-ratio testimony was judged more liberally compared to both the medium effort-ratio testimony ($p < .001, d = 1.33$) and the large effort-ratio testimony ($p < .001, d = 0.84$), while the medium effort-ratio testimony was judged more conservatively than the large effort-ratio testimony ($p = .020, d = 0.33$).

### 3.3 Discussion

As hypothesized, people in the effort-cue instruction group made more accurate judgments compared to the control group, who were not significantly more accurate than chance. Also as expected, both the medium and large effort-ratio testimony were judged more accurately compared to the small effort-ratio testimony. We also found an interaction between accuracy and testimony effort ratio. As expected, the effort-cue instruction group performed significantly better than the control group when judging the large effort-ratio testimony compared to when judging the small effort-ratio testimony, but unexpectedly there was no statistically significant difference when judging the medium effort-ratio testimony compared to when judging the large effort-ratio testimony.

The most surprising finding here is that participants performed so well on the medium effort-ratio testimony, which represented the mean effort ratio of the 22 testimonies used in Experiment 1. The performance here was expected to average somewhere between the effects of the small effort-ratio testimony and the large effort-ratio testimony. Looking at the hit- and false alarm rates for the effort-cue instruction group for this testimony (see Table 2), it is evident that the effect is driven by an exceptionally low false alarm rate at 8%, as the hit rate is actually rather low at 41%. Together with the results on response bias (see Table 2), it becomes clear that this is likely attributable to the effort-cue instruction group being more conservative in their judgments, that is, more likely to judge a statement as incorrect rather than correct. The opposite was found in the control group, who exhibited a truth bias (see Bond Jr & DePaulo, 2006; Gilbert et al., 1990; Grice, 1975), that is, a tendency to believe in people’s statements.

A notable finding is that performance by the effort-cue instruction group and the control group did not significantly differ in judgment accuracy of the testimony with the smallest effort-ratio ($d = 0.38$, in favor of the effort-cue instruction group). This is surprising, as this testimony actually had more cues in correct statements rather than correct statements, so we expected the effort-cue instruction group to perform worse than the control group. One takeaway from these results is that it lends further credence to the idea that people spontaneously rely on effort cues to some extent when judging accuracy (Brennan & Williams, 1995; Jameson et al., 1993). Another takeaway is that effort cues should be a relatively safe way to evaluate eyewitness testimonies, as even the theoretically “worst” type of eyewitness testimony for this method did not yield accuracy rates much lower than chance.

### 4 General Discussion

The main aim of the present study was to investigate whether it is possible to improve how well one can assess the accuracy in statements from sincere eyewitnesses with the help of instructions about effort cues. In Experiment 1, we compared judgment accuracy between a group instructed about effort cues and a control group using their own knowledge, and found no statistically significant effect (see Figure 1). We also compared judgment accuracy between police detectives, police students and laypersons (see Table 1), and again found no statistically significant effect. In Experiment 2, we clarified the instructions used in the first experiment and again compared judgment accuracy between an effort-cue instruction group and a control group, and this time found a substantial effect (see Figure 2).
The improved judgment accuracy by the effort-cue instruction group in Experiment 2 can likely be attributed to the clearer instructions in this second experiment ("completely rely on effort cues when judging accuracy") compared to the first experiment ("consider effort cues when judging accuracy"), as well as to using a subsample of eyewitness testimonies in Experiment 2 compared to Experiment 1.

The subsample illustrates an important point about individual differences. In Experiment 2, we chose a subsample of three testimonies from the original 22 used in Experiment 1. They were chosen to encompass the range of testimonies used in Experiment 1 in terms of effort ratio. The effort ratio was calculated by dividing the number of hedges and delays in incorrect answers with the number of hedges and delays in correct answers. In Experiment 2, participants judged a subsample consisting of the testimony with the largest and smallest effort ratio, as well as the testimony closest to the mean effort ratio. The results, as expected, showed that the largest effort-ratio testimony was judged more accurately than the smallest effort-ratio testimony. However, the medium effort-ratio testimony was essentially judged as accurate as the large effort-ratio testimony. Although this was unexpected (given the positive effort ratio-accuracy correlation obtained in Experiment 1), it can be explained in terms of individual differences. That is, despite this testimony being representative of the mean effort-ratio—which should presumably have given rise to a "medium" accuracy score—this particular witness expressed itself in such a way in which almost all incorrect answers contained hedges or pauses (as evident from the low false alarm rate, see Table 2), which in turn, resulted in high accuracy judgments. This is not something to be expected for each witness with a mid-range effort ratio, but is instead likely a result for this particular witness. To check if this was the case, we decided to compare the accuracy score for this testimony in the "effort observer"-calculations with the accuracy score for the testimony second closest to the mean effort ratio. Results showed a diminished score for the "almost medium effort-ratio testimony" (effort ratio = 2.42; \(d' = 1.02\)) compared to the medium effort-ratio testimony (effort ratio = 2.27; \(d' = 2.05\); see Table S1). Thus, although this result gives credit to the use of the effort-cue instruction to improve accuracy in this experiment, it captures a more important point: that eyewitnesses will differ in their expressions, and that it is difficult to predict what a single instance of an eyewitness testimony will look like. This is also evident from the positive testimony effort ratio-accuracy correlation in Experiment 1, as well as from the results in Experiment 2 in which only two of the three testimonies were judged above chance. This is perhaps the largest difficulty in predicting eyewitness accuracy; a law enforcement worker who is tasked with evaluating an eyewitness testimony will not be able to know just how this particular witness will express him/herself, which limits the ability for any single method to deduce testimony accuracy. This could explain the observed difficulty in judging eyewitness testimony accuracy, which is at a level only moderately above chance (Ball & O’Callaghan, 2001; Clark-Foos et al., 2015; Johnson & Suengas, 1989; Lindholm, 2005, 2008a, 2008b; Schooler et al., 1986). Indeed, like in studies on deception, there seems to be no magical cue with a "one size fits all"-property that can be used to derive truth (see DePaulo et al., 2003; Luke, 2019). This is also acknowledged in confidence-accuracy studies, in which the current prevailing view is that eyewitness confidence should only be trusted as a basis for accuracy in situations where a set of conditions have been fulfilled (see Wixted & Wells, 2017). That said, it is important to point out that in these two experiments, accuracy judgments from participants in the effort-cue instruction group were at worst equal to the control group, and substantially better when judging at least two testimonies. Thus, effort cues appear to be a relatively good basis for judging accuracy in eyewitness testimonies. Nonetheless, given that the usefulness of this method varies depending on testimony effort ratio, this method would be best used in conjunction with other evidence, as an indication of which details in a testimony to have greater belief in, rather than as a basis to decide the outcome in a trial. For example, this method could be used to give an indication of what to focus on during the investigation phase of a case, such as when compiling an offender description. A future pursuit should be to try to identify what makes a certain witness use expressions with a large or small effort ratio, as well as try to understand the stability of the effort ratio over time. If these variables can be pinned down, this method could be more readily relied on. Such a pursuit would be fruitful, as the effects for the effort-cue instruction group judging the medium and large effort-ratio testimony in Experiment 2 (\(d' = 1.63\) and \(d' = 1.67\) respectively, see Table 2) are on par with the effects obtained for experienced police detectives judging statements from transcribed eyewitness testimonies in the study by Lindholm (2008b; \(d' = 1.44\)).

To the best of our knowledge, these are the first experiments that have looked at eyewitness accuracy at a statement-level in complete testimonies. Most research on eyewitness accuracy in testimonies has focused on overall testimony credibility (Bell & Loftus, 1989; Borckardt et al., 2003; Cutler et al., 1989; Lindholm, 2008a), while a few have focused on judgments of sampled statements from a testimony (Ball & O’Callaghan, 2001; Lindholm, 2008b). Whereas overall testimony credibility judgments is limited by discrepancies within the testimony (i.e., containing both correct and incorrect details) focusing on a few select statements from a testimony risks limiting generalizability (see Yarkoni, 2020). Judging statements in complete testimonies mitigates both these issues, and provides a more comprehensive view of how people evaluate eyewitness accuracy. Likewise, a strength in Experiment 1 was that we sampled 22 different testimonies as stimulus material, rather than a single testimony.

The effort-cue method is fairly straightforward in terms of implementation, and it appears that people spontaneously already rely on effort cues when judging accuracy (Brennan & Williams, 1995; Jameson et al., 1993). The current experiments corroborate this idea. Performance for the effort-cue instruction group and control group was similar in Experiment 1, and both groups also showed a positive correlation between accuracy and testimony effort ratio, such that testimonies with more effort cues in incorrect rather than correct statements were judged more accurately. Likewise, there was a similar effect of testimony effort ratio on accuracy in Experiment 2, such that testimonies with more effort cues in incorrect rather than correct statements were largely judged more accurately. However, the results from the
“effort observer” in Experiment 1, together with the effect of effort-cue instruction in Experiment 2, indicate that a complete reliance on effort cues overall has an advantage over the “spontaneous” effort cue use, which likely involves weighing in other cues with unknown predictive validity as well. Note however, that people are not necessarily aware that they rely on these cues. When Lindholm (2008b) compared cues that high-accuracy and low-accuracy testimony judges ranked as important for their judgments, results showed that rankings were largely the same. This corroborates general findings showing that people’s introspective ability is highly limited (Nisbett & Wilson, 1977). For the interested reader, we have provided correlational data in the Supporting Information between participants’ judgment accuracy scores and their ratings of a set of cues, which show only small to non-existent effects ($r_{max} = -0.24$; see Table S2).

4.1 | Limitations

The testimonies used as stimulus material in these experiments came from a laboratory study by Gustafsson et al. (2019). In this study, participants knew they were about to watch a mock crime video, and that they would be interviewed about its contents. Moreover, the interview took place right after the end of the video, and there were no stressors or threats toward the participants. This setting does not represent the setting that many real-life eyewitnesses experience, which may limit the generalizability of our findings. For example, higher stress levels (Deffenbacher et al., 2004), poor viewing conditions (Smith et al., 2019), and a short exposure time (Memon et al., 2003) would impair encoding, which would likely result in more effortful retrieval of correct memories. Moreover, a greater retention interval between the event and the interview would increase the risk for memory distortions (for overviews, see Semmler et al., 2018; Wells et al., 2006), which could strengthen the retrieval of incorrect memories, and in turn, invalidate the use of effort to delineate accuracy. Thus, to the extent possible, this method would be best suited when evaluating a testimony obtained shortly after the viewed event, similar to Wixted and Wells’ (2017) recommendations. However, as such conditions may be exceptions, this method would be best suited in conjunction with other evidence.

Another potential limitation in the present study is that participants only judged testimonies transcribed to text, rather than the more commonly occurring oral testimonies in legal cases. However, evidence indicates that performance is actually better (Lindholm, 2008b) or at worst roughly equal (Ball & O’Callaghan, 2001) when judging transcripts rather than videotaped testimonies. It could therefore be argued that judgments of transcripts should be preferable to oral testimonies. Moreover, as the effort-cue method relies only on verbal information, it could be used in a variety of mediums: written format, audio and video.

Finally, concerning difficulty in judging testimonies, we found no evidence that police detectives were better at judging accuracy compared to police students or laypersons, as everyone’s performance averaged slightly above chance (Experiment 1). These results contrast with previous findings (Lindholm, 2008b; Lindholm et al., 1997). However, comparisons should be made with caution, as two limitations were that we had no information on the police detectives’ experience working in law-enforcement, and that the analyses were also slightly underpowered. Nonetheless, at face value, these results support the idea that judging eyewitness testimonies is a difficult task.

5 | CONCLUSION

Although no miracle cure, the take home message of these experiments is that informing people about effort cues can to some extent be used to improve judgment accuracy of eyewitness testimonies. Accuracy rates for participants relying on effort cues varied from “equal to,” to “substantially better than” participants relying on their own judgment skills. This overall modest improvement appears to stem from the fact that fact finders already spontaneously incorporate effort cues into their judgments—and, more importantly—that witnesses differ in their expressions of correct and incorrect memories, which limits the usefulness of a single cue (such as effort) to deduce accuracy. Investigating these individual differences in testimony expression and the stability of these differences within a witness is a task for future research.

ACKNOWLEDGMENTS

This research was supported by a grant from the Elisabeth and Herman Rhodin Memorial Foundation and from The Lars Hierta Memorial Foundation.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://osf.io/xnc3f/files/?view_only=37748f888f5e43c1a77c604f5431e1a6 (Experiment 1) and https://osf.io/n8t2s/?view_only=f68b9393bf5b4d459c21b5c7a41ae609 (Experiment 2).

ORCID

Philip U. Gustafsson  https://orcid.org/0000-0002-4249-5887

ENDNOTES

1 Please note that the preregistration contains two errors; (1) the power for the main analysis was incorrectly typed as 78.50% instead of 77.50% and (2) the power for the subsequent analyses were mistakenly based on $\alpha = 0.07$ instead of $\alpha = .05$ and therefore appear inflated.

2 Signal Detection provides a measure of accuracy in terms of discrimination, that is, the ability to discriminate between correct and incorrect statements. Discriminatory accuracy is measured as $d’$, which is estimated by taking the z-score of the false alarm rate (the probability that a participant judges an incorrect statement as correct) and subtracting it from the z-score of the hit rate (the probability that a participant judges a correct statement as correct). Larger positive values of $d’$
indicate a greater ability to judge accuracy whereas the reverse is true for negative values. A value of zero indicates chance performance.

3 Bias is measured as c, which is estimated by adding the standardized hit- and false alarm rate and dividing the sum by 2. Larger positive values indicate a greater tendency to respond “incorrect” (i.e., conservative bias), whereas negative values indicate a greater tendency to respond “correct” (i.e., a liberal bias). A c value of zero indicates no bias.

4 The control question was made up of a 120-word text that essentially asked if participants had read the instructions in the experiment. The text ended with a request to type in “I have read the instructions” in an empty text box if they had done so, and ignore a list of activities that were presented afterwards. The 46 excluded participants had not typed in the requested text, but had instead circled one (or several) of the activities in the list.

REFERENCES

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

Ball, C. T., & O’Callaghan, J. (2001). Judging the accuracy of children’s recall: A statement-level analysis. Journal of Experimental Psychology: Applied, 7(4), 331–345. https://doi.org/10.1037/1076-898X.7.4.331

Borckardt, J. J., Sprohge, E., & Nash, M. (2003). Effects of the inclusion and refutation of peripheral details on eyewitness credibility. Journal of Applied Social Psychology, 33(10), 2187–2197. https://doi.org/10.1111/1559-1816.2003.tb01880.x

Brennan, S. E., & Williams, M. (1995). The feeling of another’s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. Journal of Memory and Language, 34(3), 383–398. https://doi.org/10.1006/jmle.1995.1017

Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. Law and Human Behavior, 30, 31–50. https://doi.org/10.1007/s10979-006-9002-7

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indi- ces of memory processes not just markers of accuracy. Applied Cogni- tive Psychology, 22, 827–840. https://doi.org/10.1002/acp.1486

Clark-Foos, A., Brewer, G., & Marsh, R. L. (2015). Judging the reality of others’ memories. Memory, 23(3), 427–436. https://doi.org/10.1080/09658211.2014.936364

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771157

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1989). The eyewitness, the expert psychologist, and the jury. Law and Human Behavior, 13(1), 311–332. https://doi.org/10.1007/bf01067032

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McCorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. Law and Human Behavior, 28(6), 687–706. https://doi.org/10.1007/s10979-004-0565-x

DePaulo, B. M., Lindsay, J. L., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Garrett, B. (2011). Convicting the innocent. Harvard University Press.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. Journal of Personality and Social Psychology, 59(4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax & Semantics, 3. Academic Press.

Gustafsson, P. U., Lindholm, T., & Jönsson, F. U. (2019). Predicting accuracy in eyewitness testimonies with memory retrieval effort and confidence. Frontiers in Psychology, 10, 703. https://doi.org/10.3389/fpsyg.2019.00703

Hashtroudi, S., Johnson, M. K., & Chrosniak, L. D. (1990). Aging and qualitative characteristics of memories for perceived and imagined complex events. Psychology and Aging, 5(1), 119–126. https://doi.org/10.1037/0882-7974.5.1.119

Innocence Project. (2021). Innocence Project. Retrieved from https://www.innocenceproject.org/

Jameson, A., Nelson, T. O., Leonesio, R. J., & Narens, L. (1993). The feeling of another person’s knowing. Journal of Memory and Language, 28(3), 320–335. https://doi.org/10.1006/jmla.1993.1017

Johnson, M. K., & Suengas, A. G. (1989). Reality monitoring judgments of other people’s memories. Bulletin of the Psychonomic Society, 27(2), 107–110. https://doi.org/10.3758/bf03329910

Johnson, M. K., Bush, J. G., & Mitchell, K. J. (1998). Interpersonal reality monitoring: Judging the sources of other people’s memories. Social Cognition, 16(2), 199–224. https://doi.org/10.1521/soco.1998.16.2.199

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, 114, 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67–85. https://doi.org/10.1037/0033-295x.88.1.67

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowl- edge questions. Journal of Memory and Language, 32, 1–24. https://doi.org/10.1006/jmle.1993.1001

Kensinger, E. A., & Schacter, D. L. (2006). Reality monitoring and memory distortion: Effects of negative, arousing content. Memory & Cognition, 34(2), 251–260. https://doi.org/10.3758/bf03193403

Keogh, L., & Markham, R. (1998). Judgements of other people’s memory reports: Differences in reports as a function of imagery vividness. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 12(2), 159–171. https://doi.org/10.1002/sici.1099-0720(199804)12:2<159::aid-acp506>3.0.co;2-j

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

Koriat, A. (2006). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), Cambridge handbook of consciousness (pp. 289–395). Cambridge University Press.

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

Lindholm, T. (2005). Group-based biases and validity in eyewitness credi- bility judgments: Examining effects of witness ethnicity and presenta- tion modality. Journal of Applied Social Psychology, 35, 1474–1501. https://doi.org/10.1111/j.1559-1816.2005.tb02180.x

Lindholm, T. (2008a). Validity in judgments of high- and low-accurate eyewitnesses of own and other ethnic groups. Legal and Criminological Psychology, 13, 107–121. https://doi.org/10.1348/135532506X152949

Lindholm, T. (2008b). Who can judge the accuracy of eyewitness statements? A comparison of professionals and lay-persons. Applied Cognitive Psychology, 22, 1301–1314. https://doi.org/10.1002/acp.1439

Lindholm, T., Christianson, S. Å., & Karlsson, I. (1997). Police officers and civilians as witnesses: Intergroup biases and memory performance. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 11(5), 431–444. https://doi.org/10.1002/sici.1099-0720(199710)11:5<431::aid-acp470>3.0.co;2-9
