Judgment and Decision Making, Vol. 7, No. 5, September 2012, pp. 590–601

Improved realism of confidence for an episodic memory event

Sandra Buratti* Carl Martin Allwood†

Abstract

We asked whether people can make their confidence judgments more realistic (accurate) by adjusting them, with the aim of improving the relationship between the level of confidence and the correctness of the answer. This adjustment can be considered to include a so-called second-order metacognitive judgment. The participants first gave confidence judgments about their answers to questions about a video clip they had just watched. Next, they attempted to increase their accuracy by identifying confidence judgments in need of adjustment and then modifying them. The participants managed to increase their metacognitive realism, thus decreasing their absolute bias and improving their calibration, although the effects were small. We also examined the relationship between confidence judgments that were adjusted and the retrieval fluency and the phenomenological memory quality participants experienced when first answering the questions; this quality was one of either Remember (associated with concrete, vivid details) or Know (associated with a feeling of familiarity). Confidence judgments associated with low retrieval fluency and the memory quality of knowing were modified more often. In brief, our results provide evidence that people can improve the realism of their confidence judgments, mainly by decreasing their confidence for incorrect answers. Thus, this study supports the conclusion that people can perform successful second-order metacognitive judgments.

Keywords: calibration, second-order judgments, confidence judgments, metacognition, recall memory, remember/know, retrieval fluency.

1 Introduction

Realistic confidence judgments about retrieved memories are important in a number of contexts (e.g., medical and legal contexts). For example, an eyewitness to a crime must judge his or her level of confidence about correctly having identified the criminal. Although many witnesses may feel confident about their identification, the relation between identification confidence and the correctness of the identification is weak (Brewer & Wells, 2011; Sporer, Penrod, Read, & Cutler, 1995). In spite of this weakness, research has also shown that jurors often judge eyewitness credibility based on the level of confidence the eyewitness expresses (Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel,1981; Wells, Ferguson, & Lindsay, 1981). Thus, the level of confidence about a memory report should be as accurate as possible relative to the correctness of the report (Leippe & Eisenstadt, 2007).

In general, the realism of confidence judgments pertains to how well a person's confidence for a memory report matches the correctness of the report (confidence realism is also called confidence accuracy; e.g., Yates, 1994). The concept of confidence realism includes two aspects: calibration, the relationship between the level of the confidence judgments and the probability of the answer being correct; and discrimination, the extent to which the respondent can discriminate between correct and incorrect answers by means of their confidence judgments. In this study, we attended only to participants' ability to improve calibration.

Numerous studies have reported that people often show overconfidence (e.g., they are more confident than their memory report is correct). This is the case both for general knowledge questions (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994) and event memory questions (e.g., Allwood, 2010; Allwood, Ask, & Granhag, 2005; Allwood, Innes-Ker, Holmgren, & Fredin, 2008; Leippe & Eisenstadt, 2007), although the basis of this so-called overconfidence effect has been widely debated (see e.g., Brenner, Koehler, Liberman, & Tversky, 1996; Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Brenner, 2004; Koriat, 2012). Given that people show a lack of realism in their confidence judgments in many contexts, finding ways to help people improve the realism of their confidence judgments is important.

Our first aim was to investigate whether individuals can increase the realism of their confidence judgments of memory reports by adjusting confidence judgments they believe are the most unrealistic. The task of improving the realism of a confidence judgment, designated here as

This research was funded by the Swedish Research Council (VR) with a grant given to the second author.

*Department of Psychology, University of Gothenburg, Box 500, SE-40530, Göteborg, Sweden. Email: sandra.buratti@psy.gu.se.

†Department of Psychology, University of Gothenburg. Email: cma@psy.gu.se.

the Adjustment task, involves first identifying which confidence judgments to adjust and then modifying them.

When trying to improve the realism of confidence judgments, people presumably rely on various types of cues. Our second aim was to investigate the potential usefulness of two such cues in identifying confidence judgments that are in need of adjustment and modifying them: retrieval fluency (the subjective feeling of ease of recall) and the phenomenological quality of the retrieved memories, either Remember (associated with concrete, vivid detail) or Know (associated with a feeling of familiarity). Other cues, such as consistency of the activated evidence for an answer and the amount of information retrieved, are also available to participants (Koriat, 2012), but we did not investigate them here.

1.1 Regulating the realism of confidence

Many authors have tried different methods for increasing the realism of confidence, and some studies have shown that extensive feedback improves the realism of confidence (for reviews see Fischhoff, 1982, and Griffin & Brenner, 2004; for a broad introduction to debiasing, see Larrick, 2004). As far as we know, however, only one study has investigated techniques for increasing confidence realism of episodic memory information (Buratti & Allwood, 2012). Buratti and Allwood investigated the generalizability of an important principle of Koriat and Goldsmith's (1996) memory model for the regulation of the realism of confidence judgments. The principle is that people can regulate the correctness of their memory performance if they can choose which items to report. In the Buratti and Allwood study, the participants answered 50 questions about a video clip they had just seen. Immediately after each question, participants gave a confidence judgment about their answer to that question. If they did not know the answer, they were asked to guess. Participants were then asked to exclude answers they believed had the most unrealistic confidence judgments. More specifically, they were tasked with trying to improve their calibration, but they were not asked to improve their discrimination. The results showed that the participants who answered directed recall questions (for example, "What was the color of the car?" with no answer alternatives provided) increased the calibration of their report by excluding the confidence judgments they believed were the most unrealistic. Participants who answered recognition questions (i.e., with answer alternatives provided) were not able to increase their confidence realism.

In that study, however, the effect of these efforts to increase the realism of confidence judgments was small, although statistically significant. In the present study, we explored another technique for helping participants increase the realism of confidence judgments about directed recall questions. Specifically, we asked if the self-regulation principle in Koriat and Goldsmith's model, described above, generalizes to a situation of modifying rather than simply deleting confidence judgments deemed most unrealistic. The first part of the Adjustment task involved identification of the candidates for adjustment, which means evaluation of the realism of first-order confidence judgments. In the second part, participants gave a modified confidence judgment on the basis of evidence produced in the identification part or on the basis of other evidence.

Because metacognition can be defined as "any knowledge or cognitive activity that takes as its object, or regulates any aspect of any cognitive enterprise" (Flavell, Miller, & Miller, 1993, p. 150), an activity that targets regulation of a metacognitive process can be referred to as meta-metacognition. Thus, the regulation of realism of confidence (a second-order judgment) can be seen as a form of meta-metacognition. Only a few studies have investigated the accuracy of so-called second-order judgments (Dunlosky, Serra, Matvey, & Rawson, 2005; Miller & Geraci, 2011). Dunlosky et al. (2005), studying learning of paired associates, found evidence of successful second-order judgments. Miller and Geraci (2011) found that low-performing students showed higher overconfidence in their first-order prediction of their total exam performance than did high-performing students. However, the low-performing students were more accurate in their second-order judgments of their first-order predictions. If the participants in our study managed to increase the realism of their confidence judgments, that outcome would support the hypothesis that people can make accurate meta-metacognitive judgments about their first-order confidence judgments of specific memory reports.

1.2 Retrieval fluency and phenomenological quality of memory as second-order cues

People experience performance of cognitive tasks along a continuum from effortless to effortful. This so-called processing fluency, the subjective experience of the ease of processing information, serves as a cue for people in different judgment tasks (Alter & Oppenheimer, 2009). People use processing fluency as a cue to judge whether a statement is true or not, usually with fluency associated with truth and disfluency associated with untruth (Reber & Schwarz, 1999; Unkelbach, 2007). Other research has explored properties and effects of retrieval fluency, indexed as response latency, in a recognition context (Hertwig, Herzog, Schooler, & Reimer, 2008). Moreover, people report a higher level of confidence in their memory performance for semantic knowledge tasks when they experience the retrieval as more fluent (Kel-

ley & Lindsay, 1993; Koriat, 1993). Similarly, eyewitness research has found that experience of lower cognitive effort correlates positively with confidence (Robinson, Johnson, & Herndon, 1997; Robinson, Johnson, & Robertson, 2000). These studies found that fluency may serve as a cue for confidence judgments. Specifically, we investigated whether people can use retrieval fluency (experienced ease of retrieval) to increase the realism of their confidence judgments by identifying and changing those confidence judgments that they believe to be most unrealistic and therefore most in need of modification. For example, it could be beneficial to consider modifying high confidence judgments with associated retrieval disfluency and low confidence judgments with associated high retrieval fluency.

The phenomenological quality of the retrieved memory can also be a cue for improving the realism of metacognitive processes. Two memory qualities of interest are Remember, in which the memory is experienced as concrete, vivid, and detailed, and Know, in which the person has a sense of familiarity about the recalled memory (Tulving, 1985). Seemungal and Stevenage (2002) showed that Remember responses in an episodic memory task correlated more with confidence and correctness of the memory report than did Know responses. However, these results pertained only to central details (events and actions critical to the plot). In summary, both retrieval fluency and memory quality are likely to serve as cues in both the identification and the modification part of the Adjustment task.

1.3 Hypotheses

Based on the findings of Buratti and Allwood (2012), our first hypothesis was that, in a directed recall task, participants would improve the calibration of their confidence judgments by modifying those confidence judgments they believed were the most unrealistic. In contrast to the Buratti and Allwood study, we asked participants both to select the confidence judgments with the poorest realism and to modify them. This expansion is expected to aid performance because the participants are likely to engage more deeply in the task when thinking about how to modify the level of the confidence judgments.

It is possible that many participants spontaneously interpreted our request to adjust their confidence judgments as implying that they should foremost attend to incorrect answers that might have been given confidence judgments that were too high. This possibility is in line with the findings by Buratti and Allwood (2012) that the participants chose to exclude incorrect answers to a proportionally higher extent than correct answers. Given these observations, the second hypothesis was that items with lower retrieval fluency would be modified proportionally more often than items with higher retrieval fluency.

In addition, we suspect that people can use their prior experience with different kinds of memory qualities, such as the Remember and Know qualities, for evaluating the correctness of retrieved memories. In line with this reasoning and with the findings by Seemungal and Stevenage (2002) that Remember responses had a stronger association with confidence and correctness of the memory report than did Know responses, the third hypothesis predicted that answers with the memory quality Know would be modified proportionally more often than answers with the memory quality Remember.

2 Method

2.1 Participants

The final sample consisted of 200 persons (59 men and 141 women) from the Department of Psychology's participant pool. People sign up for this pool if they want to take part in the department's experiments. Three people were excluded from the final sample because they did not follow the instructions given during the experiment. Participants' ages ranged from 17 to 66 years (M = 25.9, SD = 7.0), and a majority were students at the University of Gothenburg. As reimbursement, each participant received a movie ticket worth approximately 15 US dollars.

2.2 Design

The study had a mixed 2×3 design with the within-participant variable Task (the Confidence task vs. the Adjustment task) and the between-participants variable Condition (Control vs. Fluency vs. Remember/Know). The participants were randomly assigned to the three conditions, with the exception that sex was equally distributed between conditions.

2.3 Procedure

The participants first watched a 2-minute, 20-second video clip that depicted a theft in a park. In the video clip, a passerby steals a handbag from a woman sitting on a bench while the woman is helping another passerby pick up a pile of papers he accidentally dropped. The video clip was filmed in one shot to simulate a real eyewitness-perspective, on-the-spot view of the event.

After watching the video, participants completed a filler task consisting of a digit span task. Next, the participants received a 10-minute overview covering the concept "realism of confidence". The instruction included two examples each of overconfident, perfectly realistic,

and underconfident persons. The overview included instructions for using a confidence scale.

After the instruction, the participants completed a 10-question test about the concept of realism of confidence. For each of four questions, the participants had to decide whether the assertion in the question was true or not. For example: "A person who makes on average low confidence ratings such as 0%, 20%,10%, of how confident he or she is in the information he or she reported, can still be perfectly realistic on average in their confidence judgments" (correct answer: TRUE). The remaining six questions gave examples of a person answering questions about a video clip and providing confidence judgments about their answers. The participants then decided if the person in the example was overconfident, perfectly realistic, or underconfident. In brief, the instructions gave information only about the calibration aspect of the concept of realism of confidence.

The participants then did the Confidence task. In this task, the participants answered 40 directed recall questions about the contents of the video clip (e.g., "What color was the woman's scarf?"). No answer alternatives were provided. Participants were informed that the person with the highest proportion of correct answers in each condition would receive an extra movie ticket. They were asked to answer all questions and to guess if they did not know the answer. After each question, the participants rated their confidence that their answer was correct on a confidence scale ranging from 0% ("I'm absolutely sure that my answer is incorrect") to 100% ("I'm absolutely sure that my answer is correct") with 10% increments. After the confidence judgment, participants in the control condition proceeded to the next recall question.

In the fluency condition, the confidence scale was followed by a retrieval fluency scale measuring the ease of retrieval. Participants rated how fluently the answer came to mind by answering the question, "How easy/hard was it to recall the answer?" (1 "Very hard to recall" to 7 "Very easy to recall"), before proceeding to the next recall question. In the Remember/Know condition, before proceeding to the next recall question, the participants reported the experienced memory quality of their answer (Remember, Know, or Guessing) by answering a three-alternative choice question. Following the procedure of Gardiner and Richardson-Klavehn (2000), if the participants remembered that they saw or heard the information they needed to answer the question, they were to mark the Remember box. If they instead had a feeling that they "just know the answer" because it felt familiar to them, they were to mark the Know box. Finally, if neither of these alternatives was adequate, they were to mark the Guessing box. The order of the two first boxes (Remember and Know) was counterbalanced across participants.

Participants then proceeded to the Adjustment task. In this task, they were asked to improve their confidence judgments by adjusting those judgments from the Confidence task that they believed were the least realistic. Each participant recorded the new modified confidence judgments on a text line provided next to the confidence scale. Participants could change as many confidence judgments as they wanted but a minimum of 20 changes was recommended. The participants were also told that the change should be in at least 10 percent units and that they could change only the confidence judgments, not their answers to the questions from the Confidence task. The participant in each condition with the most realistic confidence judgments after the Adjustment task was to receive an extra movie ticket.

2.4 Measures of realism of confidence

There are several measures for assessing realism of confidence. One measure is bias, in which the proportion of correct answers is subtracted from the average level of confidence (Yates, 1994). A bias value of zero indicates perfect realism in this respect, a value above zero indicates overconfidence, and a value below zero underconfidence. We also used a modified version of the bias measure, called absolute bias, which is the absolute difference between a person's proportion correct and confidence level. This version of the bias measure is similar to the version used by Bruine de Bruin, Parker, and Fischhoff (2007). The modified measure captures absolute improvements (changes towards zero).

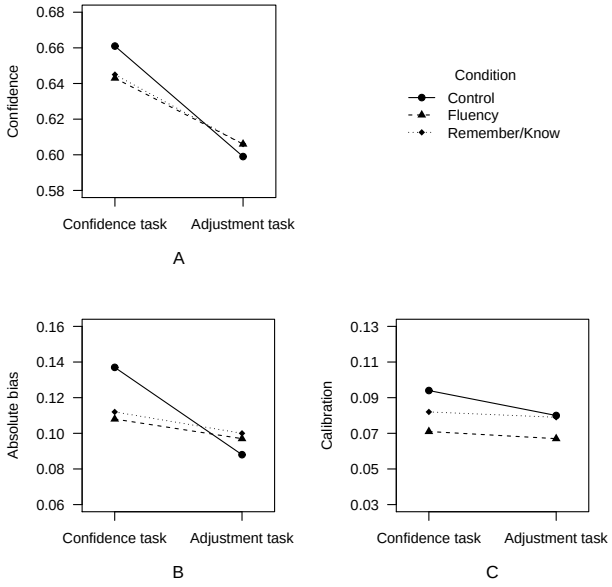
We also used the calibration measure, which is the mean squared deviation of confidence from the proportion of correct answers in each confidence class (0%, 10%, 20%, etc.), with confidence classes weighted by the number of items in each class (see Lichtenstein et al., 1982).

3 Results

The results from the knowledge test about the concept "realism of confidence" showed that 66% of the participants had all 10 answers correct, 78% had 9 or more answers correct, and 96% had 8 or more answers correct. However, the results did not differ in the following analyses with inclusion of only the participants with 10 correct answers compared to inclusion of all participants.

In the Adjustment task, participants adjusted on average 12.9 (SD = 6.6) confidence judgments (range 0–40). There was no significant difference between conditions in the number of modified confidence judgments. The participants with fewer correct responses modified more confidence judgments than those with more correct responses ($r = -.15$, $p < .030$). This result is in line with the

Figure 1: Level of confidence (A), absolute bias (B), and calibration (C) for each condition across tasks.



result that the participants with more correct responses were also more realistic in their confidence judgments as measured by absolute bias (r = −.44, p < .001) and calibration (r = −.33, p < .001). This outcome replicates the finding by Lichtenstein and Fischhoff (1997) that people who know more also seem to know more about how much they know.

3.1 Increasing the realism of confidence

In order to investigate whether the participants managed to increase their realism, the data were analyzed in 2×3 ANOVAs with the within-participant factor of Task (the Confidence task vs. the Adjustment task) and the between-participants factor of Condition (Control vs. Fluency vs. Remember/Know). To facilitate comparison, the results from the analyses of confidence, absolute bias, and calibration are presented in the panels A, B, and C of Figure 1. For calculations of effect sizes for the main and interaction effects, the generalized eta squared was used (Bakeman, 2005; Olejnik & Algina, 2003); for simple effects, partial eta squared was used. Table 1 shows the descriptive results for the dependent variables for all three conditions and the two tasks.

As Figure 1A shows, when asked to improve the realism of their confidence judgments, participants decreased their level of confidence in the Adjustment task (F(1, 197) = 129.29, p < .001, generalized η^2 = .04). A significant interaction between Task and Condition was also found (F(2, 197) = 4.17, p =.017, generalized η^2 = .01). Despite this interaction, simple analysis showed that the slopes were significant for all three conditions.

Table 1: Mean and SDs for accuracy, confidence, bias, absolute bias and calibration for the the control (n = 66), fluency (n = 67) and remember/know conditions (n = 67) and the two tasks.

	Confidence task M (SD)	Adjustment task M (SD)
Accuracy[a]		
Control	.548 (.087)	
Fluency	.571 (.092)	
Remember/Know	.568 (.110)	
Confidence		
Control	.661 (.116)	.599 (.112)
Fluency	.643 (.102)	.606 (.102)
Remember/Know	.645 (.127)	.614 (.135)
Bias		
Control	.113 (.118)	.050 (.094)
Fluency	.072 (.117)	.035 (.113)
Remember/Know	.078 (.118)	.038 (.120)
Absolut bias		
Control	.137 (.089)	.088 (.059)
Fluency	.108 (.085)	.097 (.067)
Remember/Know	.112 (.086)	.100 (.076)
Calibration		
Control	.094 (.044)	.080 (.034)
Fluency[b]	.071 (.037)	.067 (.028)
Remember/Know	.082 (.035)	.079 (.041)

[a] The accuracy measure in the Adjustment task is the same as in the Confidence task.

[b] The number of participants for the calibration measure is 66 since an outlier was excluded.

As Figures 1B and 1C show, participants actually managed to increase their realism of confidence when they modified the confidence judgments they thought were the most unrealistic. There was a significant main effect of Task for both the absolute bias measure (F(1, 197) = 3.07, p < .001, generalized η^2 = .04) and the calibration measure (F(1, 196) = 9.72, p = .003, generalized η^2 = .01). In addition, there was a significant interaction between Task and Condition for the absolute bias measure (F(2, 197) = 8.09, p < .001, generalized η^2 = .02). Moreover, for the calibration measure, the interaction was almost sig-

Table 2: Mean and SDs for confidence, bias, absolute bias and calibration for the unchosen confidence judgments, the confidence judgments chosen to be modified, and the modified confidence judgments for the chosen items.

	Unchosen conf. judg.	Chosen conf. judg.	Modified conf. judg.
	M (SD)	M (SD)	M (SD)
Confidence			
Control	.696 (.147)	.611 (.144)	.400 (.219)
Fluency	.693 (.133)	.530 (.133)	.391 (.173)
Remember/Know	.719 (.167)	.552 (.168)	.413 (.179)
Bias			
Control	.059 (.102)	.236 (.230)	.138 (.199)
Fluency	.041 (.117)	.147 (.190)	.164 (.155)
Remember/Know	.056 (.117)	.137 (.252)	.166 (.132)
Absolut bias			
Control	.094 (.070)	.265 (.193)	.138 (.119)
Fluency	.094 (.076)	.190 (.146)	.164 (.116)
Remember/Know	.096 (.080)	.221 (.182)	.166 (.132)
Calibration			
Control	.073 (.044)	.251 (.179)	.136 (.085)
Fluency	.064 (.042)	.200 (.120)	.150 (.091)
Remember/Know	.068 (.041)	.191 (.148)	.147 (.101)

Note. Conf.judg. = Confidence judgments.

nificant (F(2, 197) = 2.67, p = .072, generalized η^2 =.01). Simple effects analysis showed that only participants in the control condition managed to increase their realism of confidence on both the absolute bias measure (F(1, 197) = 41.29, p < .001, partial η^2 = .18) and the calibration measure (F(1, 197) = 13.13, p < .001, partial η^2 = .09).

As can also be seen in Figures 1B and 1C, the control condition had a higher level of absolute bias and calibration than the other two conditions in the Confidence task, but not in the Adjustment task (although the difference between conditions in the Confidence task was statistically significant only for the Calibration measure, F(2 197) = 4.14, p = .017, partial η^2 = .03). Thus, the procedures in the Fluency and Remember/Know conditions might have helped participants to be more realistic already in the Confidence task. It should be noted that the results in Figure 1 are diluted by all the confidence judgments that were not modified.

Table 3: Average level of increase and decrease in confidence judgments for correct and incorrect items in the adjustment task; average number of adjustments within parentheses.

	Correct items	Incorrect items
Increase	.218 (2.1)	.243 (2.0)
Decrease	−.285 (3.0)	−.363 (5.8)
Total	−.075 (5.1)	−.208 (7.8)

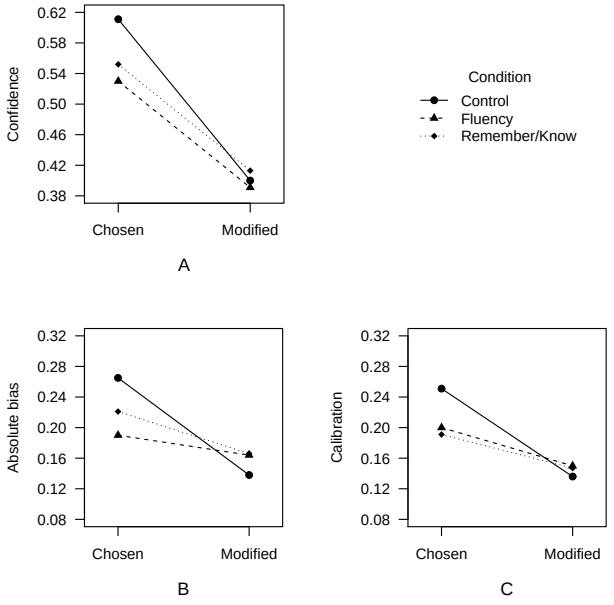
3.2 Strategies for increasing the realism of confidence

We also investigated whether the significant increase in realism as a consequence of the Adjustment task resulted from the use of a simple heuristic in which participants merely lowered their confidence judgments randomly. Our results instead showed that, to some extent, participants could identify the least realistic confidence judgments and modify them, thereby increasing their realism. Specifically, to analyze differences in confidence, absolute bias, and calibration between confidence judgments, we calculated these measures for the original confidence judgments that were unchosen for modification, for the original confidence judgments that were chosen, and for the modified new confidence judgments (Table 2). We evaluated the differences in level of confidence, absolute bias, and calibration among the three types of confidence judgments in 3×3 mixed measures ANOVAs[1] with the between-participant variable Condition (Control vs. Fluency vs. Remember/Know) and the within-participant variable Rating sets (original unchosen vs. original chosen vs. modified confidence judgments). Because the assumption of sphericity was not met, the Greenhouse-Geisser estimates are presented for all three measures. The results showed a main difference in confidence for Rating sets (F(1.92, 373.25) = 179,852, p < .001, generalized η^2 = .36). Post hoc Bonferroni analysis showed that the unchosen confidence judgments had a higher confidence level than the confidence judgments chosen to be modified. Also, participants lowered their confidence for their chosen answers, since the modified confidence judgments had a significantly lower confidence level than the chosen judgments before modification.

The results showed a difference in realism for Rating sets as measured by the absolute bias measure (F(1.70,

[1]The means of the confidence, bias, absolute bias, and calibration measures were calculated for each participant separately on an unequal number of observations because the number of confidence judgments modified varied between participants. Thus, the dependent measures (confidence, absolute bias, and calibration) used in the dependent ANOVAs are a bit less robust than is desirable.

Figure 2: Level of confidence (A), absolute bias (B), and calibration (C) for the confidence judgments chosen to be modified and the modified confidence judgments.



33.19) = 56.04, p < .001, generalized η^2 = .22) as well as by the calibration measure (F(1.52, 293.88) = 97.49, p < .001, generalized η^2 = .30). Post hoc Bonferroni analysis for both measures showed that the chosen confidence judgments had a worse level of realism than the confidence judgments unchosen to be modified. Further, the participants significantly improved the realism for the chosen confidence judgments when modifying them (p < .001). There was no main effect of Condition.

In Figure 2, the slopes between confidence judgments chosen to be modified and the modified confidence judgments can be seen for confidence, absolute bias, and calibration. Note that the scaling in Figure 2 is different from that in Figure 1, with the slopes in Figure 2 being steeper than the corresponding slopes in Figure 1. All the slopes in Figures 2A, 2B, and 2C, except for the absolute bias measure in the remember/know condition, are significant.

Participants more often decreased their confidence for incorrect items (i.e., lowered their confidence for overconfident items) than they increased confidence for correct items (i.e., increased their confidence for underconfident items). As Table 3 shows, on average, participants adjusted the confidence level for more incorrect items (7.8) than correct items (5.1). Moreover, on average, participants decreased their confidence for incorrect items by a greater magnitude (−.208), compared with the average adjustment for correct items. In fact, instead of increasing their confidence for correct answers, participants slightly decreased their confidence on average (−.075).

Figure 3: Calibration curves for the control condition for the Confidence task and the Adjustment task for (A) all confidence judgments in both tasks and (B) the confidence judgments in the Confidence task that were chosen to be modified and the modified confidence judgments in the Adjustment task.

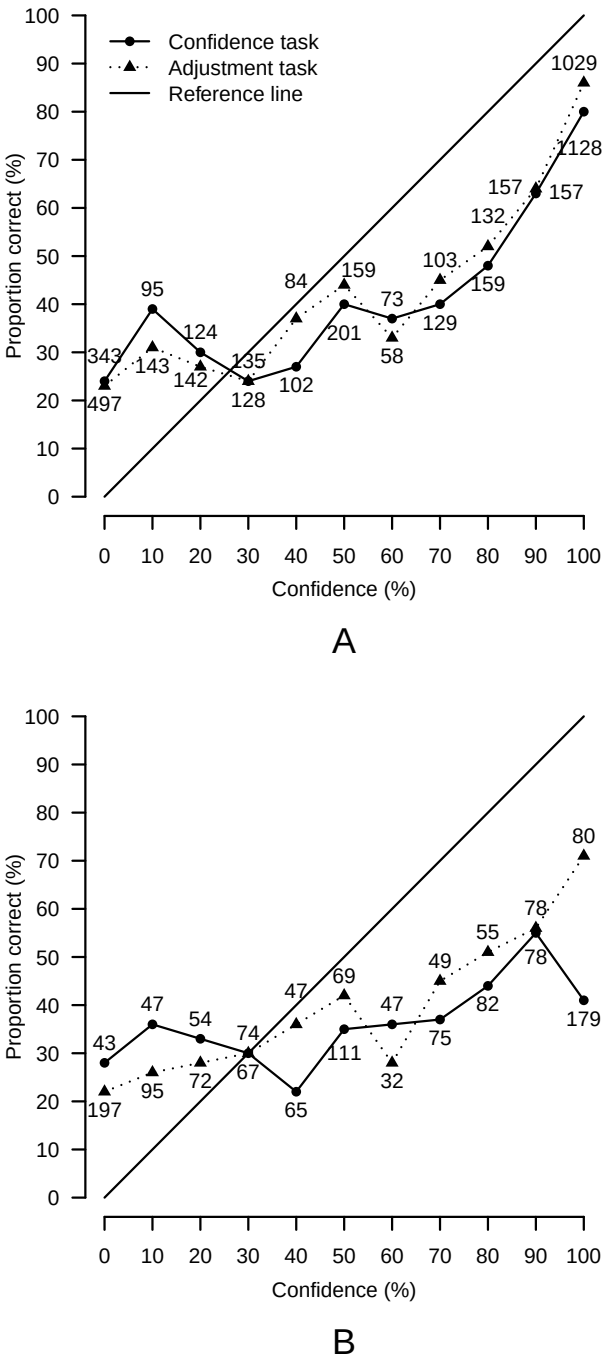


Table 4: Mean values and SDs for Remember, Know, and Guessing answers.

	Remember M (SD)	Know M (SD)	Guessing M (SD)
Bias	.120 (.122)	.206 (.215)	−.104 (.188)
Absolute bias	.131 (.111)	.238 (.180)	.165 (.137)
Calibration	.054 (.054)	.188 (.144)	.120 (.082)

3.3 Calibration curves

Because the main analysis showed that only participants in the control condition managed to increase their realism of confidence significantly in the Adjustment task, Figure 3A presents only calibration curves for the control condition, with separate curves for the Confidence task and the Adjustment task. The reference line shows perfect realism, values below the line indicate overconfidence, and values above the line indicate underconfidence. Each confidence class for each task is tagged with the number of times the confidence class was used in the task. The curves show that the realism in the Adjustment task, as measured by the distance of the points from the reference line, differs from that of the Confidence task at almost every confidence level. However, all the confidence judgments that were not modified diluted these results, and the differences tended to be small. In Figure 3B, the calibration curves are displayed for only the confidence judgments that were chosen to be modified for the control condition and the modified confidence judgments in the Adjustment task for the same answers. In this figure, the successful performance of the second part of the Adjustment task, that is, the modification of the chosen confidence judgments, is more obvious.

3.4 Fluency and confidence

Part of our second aim was to investigate participants' use of retrieval fluency as a cue when modifying confidence judgments. For the fluency condition, we found that items chosen to be modified had a lower retrieval fluency score and confidence level than unchosen items in the Confidence task. Specifically, the retrieval fluency score for the confidence judgments chosen to be modified (M = 3.10, SD = .804) was significantly lower than that for the unchosen confidence judgments (M = 4.51, SD = .808; F(1, 65) = 13.89, p < .001, generalized η^2 = .43). The confidence level was also significantly lower for the confidence judgments chosen to be modified (M = .530, SD = .133) than for the the unchosen confidence judgments (M = .692, SD = .130; F(1, 65) = 61.65, p < .001, generalized η^2 = .43).

Higher confidence judgments were associated with higher retrieval fluency in the Confidence task. Thus, the average Pearson's correlation between the fluency scores and the confidence scores for the participants showed a high value (r = .90). However, after the Adjustment task, the average correlation between the confidence judgments and fluency scores was only moderate (r = .35; p < .001). This may be expected since only some of the confidence judgments were adjusted while the original fluency scores remained the same, thus the correlation pattern is likely to have become weaker since the fairly regular association pattern is likely to have become more diversified when some of the values changed and others remained the same.

3.5 The realism of confidence for Remember/Know answers

Another part of our second aim was to investigate how the memory quality of the answer served as a cue when deciding which confidence judgments to modify. The results showed that Know answers were chosen to be modified more often than Remember and Guessing answers. On average, 20% of confidence judgments for Remember answers were modified, while 52% of confidence judgments for Know answers were modified and 43% of confidence judgments for Guessing answers were modified.

As Table 4 indicates, answers rated as Guessing showed underconfidence, whereas answers rated with the memory qualities Remember or Know showed overconfidence. Also, there was a significant difference in realism between the memory qualities when measured with both the absolute bias measure (F(2, 132) = 9.61, p = .001, generalized η^2 = .11) and the calibration measure (F(1.63, 107.53) = 29.38, p = .001, generalized η^2 = .28). Bonferroni post hoc analyses showed that Know answers were less realistic than Remember (both measures, p < .001) and Guessing answers (absolute bias, p = .028; calibration, p = .004). However, only the calibration measure showed that the Remember answers were more realistic than the Guessing answers (p < .001).

In summary, participants were most likely to adjust their confidence level for answers when retrieval was disfluent and when the answers where not associated with the memory quality Remember. Furthermore, the participants managed to increase the realism of confidence for the selected answers by adjusting their confidence levels.

4 Discussion

We investigated whether people can increase the realism of their confidence judgments for an episodic memory task when they are instructed to adjust the confidence

judgments they believe are the most unrealistic. The instruction on confidence realism given to the participants dealt only with the calibration aspect (i.e., the direct relationship between confidence and correctness), and this aspect was therefore what the participants attempted to improve. The results support our first hypothesis that people can perform this task successfully. Participants' success in improving the realism of their confidence judgments was especially clear when considering only those confidence judgments that were selected for adjustment.

In general, however, the accomplished increase in realism was small. Moreover, when we conducted simple effects analyses for the different conditions, the effect was found only for the control condition. As noted earlier, however, these modest overall effects reflect the fact that many confidence ratings were left unchanged. One reason why the increase in realism was found only in the control condition may be that the additional ratings in the fluency and Remember/Know conditions induced participants to make more stringent judgments during the initial Confidence task. Thus, the participants' better confidence realism in the two non-control conditions made it harder for them to improve their level of realism in the Adjustment task, as compared with the Control condition.

An interesting question is why participants initially showed better realism in the fluency and Remember/Know conditions. We speculate that the fluency and Remember/Know ratings made participants more sensitive to effective cues regarding the realism of their confidence judgments (i.e., degree of fluency and type of memory quality in terms of Remember/Know). For example, in the context of episodic memory, a vague feeling of knowing the answer may not be as convincing as a clear experience of remembering the answer. In addition, the task to make the fluency ratings or the remember/know ratings may have made the participants more aware of the possibility of evaluating their performance and thus of the importance of being able to justify the level of their confidence judgments. This may have increased their feeling of being accountable for their confidence ratings, maybe to others, but at least to themselves. Prior research has shown that, in situations similar to that in the present research, when the participants feel accountable for the level of their confidence judgments they may spend longer time on the task (Arkes, Christensen, Lai & Blumer, 1987), become more aware of possible counterarguments and, in general, achieve better realism in their confidence judgments (Arkes et al., 1987; Lerner & Tetlock, 1999).

It is possible that instructing participants to attend to retrieval fluency and the Remember/Know memory quality of their answers while performing confidence judgments might be a successful debiasing method *per se*. Such a debiasing strategy would be fairly simple to implement and, both for this reason and for theoretical reasons, it would be of interest to further develop and test this strategy in future research.

Turning back to the performance in the Adjustment task, the results also indicated that the improvement in realism did not arise from the use of a simple heuristic in which participants merely lowered their average confidence level for random items. Instead, participants seem to have targeted the confidence judgments with the worst levels of absolute bias and calibration and then increased the realism of these confidence judgments by altering them. Also, participants lowered their confidence for incorrect answers to a greater extent (thereby increasing their realism for overconfident items) than they raised their confidence for correct answers (that is, increased their realism for underconfident items).

Our results lend some support to the idea that the self-regulation principle of Koriat and Goldsmith's (1996) model can be generalized to the regulation of the realism of confidence judgments. However, a difference in the regulation task in the two contexts should be noted. In Koriat and Goldsmith's approach, participants were free to regulate their memory reports by choosing whether to report a memory or not. In the current study, the participants were asked to regulate their confidence judgments by modifying them. They did not have the option of regulating the realism of confidence by simply excluding problematic items.

The results also confirm our second hypothesis that participants would choose to modify confidence judgments with lower retrieval fluency scores. The retrieval fluency and confidence judgments were highly correlated. Earlier studies have shown that retrieval fluency is used as a cue when making confidence judgments (Kelley & Lindsay, 1993; Koriat, 1993). Our findings add to these studies by indicating that retrieval fluency may also be used as a cue in second-order judgments for increasing the realism of confidence. If so, this association may help to explain why proportionally more incorrect than correct items were modified.

The outcomes in this study also support our third hypothesis by showing that the participants chose more often to modify confidence judgments of items categorized with the memory quality Know. The level of realism for Remember items was significantly higher than that for Know items on the absolute bias and calibration measures and significantly more realistic than Guessing items on the calibration measure. These results are in line with those reported by Seemungal and Stevenage (2002), although our study also used realism of confidence measures that provided information about the direction of the bias in confidence judgments. Whether or not memory quality served as an explicit cue for identifying the items to be adjusted, it was a valid cue for this purpose because

the Know category was the category with the worst realism.

Two further observations are relevant here. The first is the relationship of our study with the dialectical bootstrapping method (DBM) reported by Herzog and Hertwig (2009). Herzog and Hertwig suggested DBM as a way to improve answers to questions asking for numerical estimates and predictions and this method bears similarities to the Adjustment task used in this study. The rationale for DBM is that combining somewhat different estimates will tend to cancel out many types of error, and DBM draws on previous research showing that the average of predictions from many individuals is better than a typical estimate in the same group. In the DBM, a participant first must answer a numerical estimation question and then after some time provide a new estimation (with the intent that somewhat different knowledge may be activated the second time). The final answer given in the DBM is the average between the two estimates. The DBM and our Adjustment task are similar in that they require two responses that to some extent draw on somewhat different knowledge. However, the DBM and our Adjustment task differ in two important ways. First, the Adjustment task includes identifying the confidence judgments to adjust rather than considering an answer to a knowledge task. Second, the DBM in the final step crucially involves averaging the two estimates. It is possible that some participants in the Adjustment task may have determined a final modified confidence judgment by averaging their confidence judgment from the Confidence task with an initial confidence judgment in the Adjustment task. However, we do not know if any participants used this tactic, and this issue may be investigated in future research. Such research could also include instructions urging the participants to consider new knowledge in the Adjustment task; that is, knowledge that they had not already attended to in the Confidence task. In general, future research should attempt to identify effective ways of instructing participants to improve their performance in the Adjustment task, such as by paying more equal attention to low and high confidence judgments.

The second observation is that the increase in realism of confidence could arguably have been a mere consequence of participants' making their confidence judgments twice for the same task. Allwood, Granhag, and Johansson (2003) found, however, that twice answering episodic memory questions and giving confidence judgments about these answers had very small effects on memory and metamemory performance; thus, this suggestion does not seem very likely. Similarly, Herzog and Hertwig (2009) reported that simply answering numerical estimation questions a second time (without using the DBM) did not markedly improve performance. Herzog and Hertwig concluded that the knowledge attended to on the two occasions should differ at least somewhat. We suspect that this conclusion also pertains to the successful adjustment of confidence judgments.

In conclusion, participants could use second-order metacognitive judgments to regulate the realism of their first-order confidence judgments about their answers to directed recall questions covering a filmed crime event. Our results provide a new example of a metacognitive task and indicate, in line with earlier research, that people can make accurate meta-metacognitive judgments (second-order judgments) about their first-order metacognitive performance (Dunlosky et al, 2005; Miller & Geraci, 2011). In addition to our above suggestions for further research, future investigations in this area should also try the debiasing technique investigated in this study on semantic rather than episodic memory reports.

References

Allwood, C. M. (2010). Eyewitness confidence. In P. A. Granhag (Ed.), *Forensic psychology in context: Nordic and international approaches* (pp. 281–303). Devon, UK: Willan Publishing.

Allwood, C. M., Ask, K., & Granhag, P. A. (2005). The cognitive interview: Effects on the realism in witnesses' confidence in their free recall. *Psychology, Crime & Law, 11*, 183–198. http://dx.doi.org/10.1080/10683160512331329943

Allwood, C.M, Granhag, P.A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgments: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545–561. http://dx.doi.org/10.1002/acp.888

Allwood, C. M., Innes-Ker, A. H., Holmgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime & Law, 15*, 529–547. http://dx.doi.org/10.1080/10683160802636709

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235. http://dx.doi.org/10.1177/1088868309341564

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133–144.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379–384. http://dx.doi.org/10.3758/BF03192707

Brenner, L.A., Koehler, D.J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and fre-

quency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65*, 212–219. http://dx.doi.org/10.1006/obhd.1996.0021

Brewer, N., & Wells, G. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20*, 24–27. http://dx.doi.org/10.1177/0963721410389169

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938–956. http://dx.doi.org/10.1037/0022--3514.92.5.938

Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments — Regulating the realism of confidence. *Cognitive Processing, 13,* 243–253. http://dx.doi.org/10.1007/s10339-012-0440-5

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*, 41–55. http://dx.doi.org/10.1007/bf01064273

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology, 132,* 335–346. http://dx.doi.org/10.3200/GENP.132.4.335--346

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous overconfidence and underconfidence - the role of error in judgment processes. *Psychological Review, 101*, 519–527. http://dx.doi.org/10.1037/0033--295X.101.3.519

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). New York: Cambridge University Press.

Flavell, J. H., Miller, S. A., & Miller, P. H. (1993). *Cognitive development* (3rd. ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. Craik (Eds.), *The Oxford handbook of memory* (pp. 229–244). New York: Oxford University Press.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models - a Brunswikian theory of confidence. *Psychological Review, 98*, 506–528. http://dx.doi.org/10.1037/0033-295X.98.4.506

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–198). Malden: Blackwell Publishing.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1191–1206. http://dx.doi.org/10.1037/a0013025

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind. Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*, 231–237. http://dx.doi.org/10.1111/j.1467-9280.2009.02271.x

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

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639. http://dx.doi.org/10.1037/0033-295x.100.4.609

Koriat, A. (2012). The self-consistence model of subjective confidence. *Psychological Review*, *119*, 80–113. http://dx.doi.org/10.1037/a0022171

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517. http://dx.doi.org/10.1037/0033-295X.103.3.490

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Malden: Blackwell Publishing.

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R. C. L. Lindsay, D. F. Ross, J. D. Read & M. P. Toglia (Eds.), *Handbook of eyewitness psychology* (Vol. 2, pp. 377–425). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255–275.

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159–183. http://dx.doi.org/10.1016/0030-5073(77)90001--0

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 79–89. http://dx.doi.org/10.1037/0021-9010.66.1.79

McClelland, A.G.R. & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models

1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Oxford, England: John Wiley & Sons.

Miller, T. M., & Geraci, L. (2011). Unskilled but unaware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 502–506. http://dx.doi.org/10.1037/a0021802

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*, 434–447. http://dx.doi.org/10.1037/1082-989x.8.4.434

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition: An International Journal, 8*, 338–342. http://dx.doi.org/10.1006/ccog.1999.0386

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416–425. http://dx.doi.org/10.1037/0021-9010.82.3.416

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied, 6*, 207–221. http://dx.doi.org/10.1037//1076-898x.6.3.207

Seemungal, F. V., & Stevenage, S. V. (2002). Using state of awareness judgements to improve eyewitness confidence-accuracy judgements. In P. Chambres, M. Izaute & P.-J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 219–231). Dordrecht, Netherlands: Kluwer Academic Publishers.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A metaanalysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*, 1–12. http://dx.doi.org/10.1037/h0080017

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 219–230. http://dx.doi.org/10.1037/0278-7393.33.1.219

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*, 688–696. http://dx.doi.org/10.1037/0021-9010.66.6.688

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). Oxford, England: John Wiley & Sons.