How Do We Judge Our Confidence? Differential effects of meta-memory feedback on eyewitness accuracy and confidence

Journal:	*Applied Cognitive Psychology*
Manuscript ID	ACP-19-0127.R1
Wiley - Manuscript type:	Research Article
Keywords:	meta-memory, feedback, eyewitness testimony, eyewitness confidence

How Do We Judge Our Confidence? Differential effects of meta-memory feedback on eyewitness accuracy and confidence

Abstract

According to the Cue-Belief Model, we assess confidence in our memories using self-credibility cues that reflect beliefs about our memory faculties. We tested the influence of meta-memory feedback on self-credibility cues in the context of eyewitness testimony, when feedback was provided prior to "testifying" via a memory questionnaire (Experiment 1) and after an initial memory questionnaire but before participants had to retake it (Experiment 2). Participants received feedback (good score, bad score, or none) on a fictitious scale purported to predict eyewitness memory ability. Those given good score feedback before testifying were more confident (but no more accurate) than those given bad score feedback. Feedback also affected confidence (good increased, bad decreased) and accuracy (good increased) *after* testifying, but only on leading questions. These differential effects of meta-memory feedback on confidence for normal and leading questions aren't explained by the Cue-Belief model. Implications for our confidence judgments are discussed.

Keywords: meta-memory, feedback, eyewitness testimony, eyewitness confidence

Introduction

Generally, people tend to believe that eyewitness testimony is reliable when the eyewitness testifies with high confidence. Mock jury studies of eyewitness credibility reveal that participants are more likely to view highly confident testimonies as more reliable (e.g., Bradfield & Wells, 2000; Sauer, Palmer, & Brewer, 2017). Until recently, most of the research on the Confidence-Accuracy (CA) relationship suggested that eyewitness confidence is *not* a good predictor of testimonial accuracy (Eisenstadt & Leippe, 2009; Krug, 2007), though there were some studies that found a strong relationship between confidence and accuracy (Lindsay, Read, & Sharma, 1998; Wixted & Wells, 2014). According to these studies, and contrary to the general beliefs of jurors and judges, we should be careful in using testimonial confidence as a marker for testimonial accuracy.

One of the reasons for the poor CA relationships observed in the literature (aside from potential measurement issues; Wixted & Wells, 2017) is that eyewitness testimonial confidence does not merely reflect memory accuracy—confidence judgments are also affected by subjective evaluations of memory accuracy. Leippe and colleagues (2009) explained confidence judgment processes, and the subjective factors that affect them, using the *Cue-Belief model*. According to this model, our confidence judgments are a product of two decision processes. First, we sense the accessibility of the memory trace elicited by a stimulus (i.e., the degree of familiarity elicited by the stimulus). Second, we determine the likelihood that the memory trace is accurate based on three types of cues. *Intrinsic cues* provide knowledge about the types of memories that are accurate (e.g., how much of the stimulus comes to mind readily, the vividness of the memory). *Self-credibility cues* inform our beliefs about our own memory capabilities (e.g., "How good am I at memorizing and remembering people's faces?"). *Extrinsic cues* relate to factors in the witnessing situation (e.g., recalling that the target was observed at a close distance) or the testing situation (e.g., cues provided by the lineup instructions and/or investigator). These cues can be affected by both internal (i.e., self-generated) and external (e.g., investigator feedback) sources. Based on these different cues, we determine the likelihood that our memory is accurate and produce a corresponding confidence judgment.

Although there is a considerable amount of research on intrinsic and extrinsic cues (e.g., biased instructions), few studies have examined self-credibility cues (Leippe et al., 2009). Leippe and colleagues (2006; 2009) manipulated scores on a memory questionnaire for a mock crime video and then had participants make lineup identifications. Participant self-credibility cues were manipulated in the form of

feedback about questionnaire scores. Participants who received "good score" feedback (i.e., were told that they scored in the 91st percentile on a memory test related to the event) were more confident in their identifications, and participants who received "bad score" feedback (i.e., 21st percentile) were less confident in their identifications (with participants who did not receive feedback falling in the middle).

Our primary objective was to provide a more in-depth examination of the processes underlying our confidence judgments. Using the Cue-Belief model, we examined the influence of self-credibility cues, which have been previously under-researched. One might argue that it is unlikely that real-world eyewitnesses receive explicit self-credibility cues. Though we agree that (ideally) eyewitnesses rarely receive explicit feedback about their memory abilities in real life, they almost certainly hold general beliefs about their own memory abilities, beliefs which can affect confidence in memory processes (Leippe et al., 2009). These beliefs may be sensitive to several factors, including social pressure and investigator behavior. For instance, eyewitnesses (and their accuracy) may come under intense and stressful scrutiny from the public, investigators, and the defense/prosecution. When an eyewitness testifies, their accuracy is of central interest. The heightened state of stress and scrutiny may make eyewitnesses especially cognizant (and potentially skeptical) of their own memory abilities, which in concert with the desire to give the impression of reliability could leave them particularly susceptible to feedback (Roper & Shewan, 2002). Thus, it is not unreasonable to expect that eyewitnesses engage in deeper consideration of their memory abilities and/or receive at least *implicit* feedback about these abilities. A better experimental understanding of self-credibility cues can provide insight into the mechanisms underlying real-world eyewitness memory processes.

Across two experiments, we manipulated self-credibility cues by giving good, bad, or no feedback to participants on a fictitious assessment scale that ostensibly measured eyewitness memory abilities. We then examined whether this manipulation affected eyewitness testimonial confidence and accuracy. Before administering the assessment scale, the experimenter explained to participant-eyewitness that it could predict the accuracy of eyewitness testimony. The experimenter explained that a good score indicated accurate eyewitness memory, while a bad score indicated inaccurate memory. We assumed that feedback would influence self-credibility cues, which themselves reflect beliefs about our own memory faculties. If we generate confidence judgments in accordance with the Cue-Belief model, participants receiving good feedback should report high confidence while participants receiving bad feedback should report low confidence.

Our study was unique in that we manipulated self-credibility cues more generally by using a fictitious assessment scale that was not directly related to the event witnessed by participants. In previous studies (e.g., Leippe et al., 2006; 2009), the fictitious scale was directly related to the mock crime video that participants watched (i.e., it only included questions specifically about the video). In contrast, our scale was designed to manipulate beliefs about participants' memory faculties in general. Though event-specific memory feedback (e.g., "Are you sure you were able to see the man's face?") may be more common in the real-world than explicit or implicit general memory feedback (e.g., "Are you sure you are able to remember faces well?"), it is not unreasonable to expect that both types of feedback influence eyewitnesses' perceptions of their own memory abilities—in particular during eyewitness interviewing where there is great social pressure and eyewitnesses are likely to be susceptible of any kind of investigator feedback (Roper & Shewan, 2002). Additionally, the Cue-Belief framework has been applied to identification decisions (Leippe et al., 2006; 2009), but the effects of self-credibility cues on more general types of questions (that are common during investigations) have not been tested.

We had two other goals with this study. First, we tested whether the influence of investigator feedback varied by question type (normal vs. leading questions). In our study, leading questions implied the existence of something that was not actually in the previously witnessed event. Investigators may use leading questions in order to make up for a lack of information and confirm their own inferences about a case. There are few studies that have examined the interaction between question type and investigator feedback. Previous studies investigating this relationship have adopted the Selective Cue Integration Framework (SCIF; Charman, Carlucci, Vallano, & Gregory, 2010). According to SCIF, we assess our confidence using internal cues, namely, memory traces and *ecphoric similarity*, the degree of perceived similarity between the stimulus and its memory trace (Bradfield, Wells, & Olson, 2002). When internal cues are weak (memory traces are not highly accessible, or ecphoric similarity is low), we search for external cues that we believe are reliable and take them into account. Thus, in the context of a *question type* by *investigator feedback* interaction, SCIF predicts an effect of feedback on leading questions (where memory traces do not exist, and if they do, ecphoric similarity is low) but not on normal questions (where stronger memory traces should exist and ecphoric similarity is more likely to be high). For instance, consider the leading question "What kind of gun did the culprit have?", where the presence of a weapon is not certain. The eyewitness may indeed have internal cues relevant to the question—specifically its negation (e.g., A memory of a knife, or of the culprit's empty

hand). However, these memory traces will have low ecphoric similarity to the stimulus (i.e., the gun) and thus, confidence may be lower. In our study, we explored the effect of feedback on both normal and leading questions to gain more insight into the potential interaction between question type and investigator feedback, and to test the validity of predictions made by both the Cue-Belief model and the SCIF.

Second, we explored whether administrator feedback effects are affected by the *timing* of feedback. In our first experiment, participants were given feedback before "testifying" (i.e., answering the event memory questionnaire). In our second experiment, participants were given feedback after testifying (on an initial event memory questionnaire) and were then required to testify again (repeat the same memory questionnaire). Many prior researchers have examined the effects of feedback before and after testifying (e.g., Steblay, 1997), but pre- and post-testimony feedback effects are usually examined separately. However, in the context of self-credibility cues, only the effects of pre-testimony feedback have been examined (Leippe et al., 2009). To explore whether timing is relevant to the Cue-Belief model, we tested the effects of self-credibility feedback both before testimony (Experiment 1) and after testimony (Experiment 2).

Experiment 1

In Experiment 1, we focused on the influence of pre-identification feedback on eyewitnesses' self-credibility cues. Although few studies have examined the effects of pre-identification feedback, Leippe and colleagues (2006) did test whether pre-identification feedback affected eyewitnesses' identifications. However, they employed a manipulation directly related to the specific witnessed event (i.e., score feedback was purportedly related to memory performance on questions about the witnessed event). In our study, we used a fictitious scale that included items related to participants' general memorial and cognitive faculties. In addition, instead of making an identification, participants in our study answered normal and leading questions about the witnessed event, with question accuracy and confidence serving as dependent variables. If we judge confidence using self-credibility cues as per the Cue-Belief model, our feedback manipulation should have similar effects to those observed in Leippe and colleagues work (2009). Furthermore, according to SCIF, feedback should affect confidence on leading questions (where internal cues are not available and participants must rely on external cues) but not on normal questions (where strong internal cues are available). Specifically, we predicted the following: good score feedback confidence = no score feedback confidence = bad score feedback confidence for normal questions, and good

score feedback confidence > no score feedback confidence > bad score feedback confidence for leading questions. Because the Cue-Belief model implies that feedback affects subjective assessments of accuracy rather than memory accuracy itself, we did not predict any effects of feedback on accuracy (i.e., good score feedback accuracy = no score feedback accuracy = bad score feedback accuracy).

Method

Participants and Design. Participants in Experiment 1 included 90 Japanese students from the University of Tokyo (37 males, 53 females; Age $M = 19.72$, $SD = .88$). Our design was a 3 (Feedback condition: good score, bad score, no feedback) x 2 (Question type: Normal, Leading) mixed design, with Feedback condition as a between-subjects IV, Question type as a within-subjects IV, and accuracy and confidence as separate DVs. Thirty participants were randomly assigned to each of the three feedback conditions (good score, bad score, no feedback). The good and bad score conditions were treated as experimental conditions, with the no feedback condition serving as the control condition. All participants received compensation worth 500 yen.

Materials and Procedure. First, the experimenter stated the purpose of the experiment, giving the cover story that the study was examining the validity of a (fictitious) assessment scale designed to predict eyewitness testimonial accuracy. Next, the participants completed the fictitious assessment scale and watched the mock crime video. After the video, participants received feedback about their score on the assessment scale. Participants were randomly given one of the three feedback types (good score, bad score, no feedback). The experimenter was aware of the assigned condition before the experiment began. After receiving good score feedback, bad score feedback, or no feedback, participants completed the questionnaire about the mock crime video. Participants were instructed to complete all questions but were told that sometimes the correct answer to a question was "None" (i.e., the queried event or item did not appear in the video). All participants completed the experiment individually. Finally, the experimenter asked participants if there was anything they noticed about the experiment. We did not specifically ask if they were suspicious about the experimental manipulation. However, we believe that the fact that *no* participants mentioned suspicion renders it unlikely that a substantial proportion of the sample were wise to our manipulation. In their experiment, Leippe and colleagues (2009) reported that 7.7% of their sample were suspicious of the manipulation. If we assume that the proportion in our sample was similar or smaller, it is unlikely that a subsample of suspicious

participants affected our results.

After answering this question, participants were fully debriefed. In the debriefing, participants were informed that the assessment scale they completed was fictitious and were asked whether they had suspected that it was. No participants indicated via the final question or during debriefing that they had suspected that the assessment scale was fictitious.

Fictitious assessment scale. The fictitious scale consisted of three sections. Section 1 consisted of items about everyday cognition (e.g., "When I go to a new place for the first time, I will later be able to recognize the people I meet there") while the items in Section 2 concerned the type of cognition (e.g., "I focus on details first"). Section 3 consisted of items about actions and behaviors (e.g., "When I do something, I will spend a lot of time completing it"). Each section included seven items, all scored on 8-point Likert scales (0 = *Entirely disagree*; 7 = *Entirely agree*). All items were generally related to their Section topic but were arbitrarily created for this experiment. Total scores on the three sections were described using grades: A, B, or C (though no scores of B were actually given). The experimenter explained to the participants that "this assessment scale will predict your eyewitness testimonial accuracy. If your total score is an A, your eyewitness testimony tends to be extremely accurate. If your total score is a C, that means that your eyewitness testimony tends not to be accurate. If your total score is a B, that means your eyewitness testimony is of average accuracy. Please complete the assessment scale." After watching the crime video, participants received their "assessments", with the "grade" (A or C in the "good feedback" and "bad feedback" conditions respectively) written on them. Participants in the "no feedback" condition received neither their assessment scale or feedback (they were told "We are still marking your assessment scale, and you will receive it at the end of the experiment").

Mock crime video. Participants watched a 1-minute mock crime video depicting a luggage thief. In the video, two men enter a racecourse and place their luggage on their seats. Then they leave their seats, leaving the luggage behind. After a while, another man approaches and sits in the row behind. After glancing around, the man picks up one of the bags and walks away. The video was recorded obliquely from behind. The thief was on the screen for approximately 20 seconds and in frontal view for about 5 seconds.

Mock crime video questionnaire. The questionnaire consisted of 10 recall questions, three of which were leading questions (see Table 1). In addition to answering each question, participants rated their confidence in each answer on a 5-point Likert scale (0 = *Not at all confident*; 4 = *Very confident*). Accuracy and confidence on this questionnaire (averaged separately across 7 normal questions and 3 leading questions) were our primary dependent variables. Similar to the original study and other investigator feedback studies (Leippe et al., 2009; Quinlivan, Neuschatz, Douglass, Wells, & Wetmore, 2012), participants were required to provide an answer to all recall questions—there was no option to say "I don't know/I am not sure". See Supplementary Material A for scale reliabilities.

Results

First, we conducted a two-way mixed ANOVA with feedback condition (good score, bad score, no feedback; between-subjects) and question type (normal, leading; within-subjects) as the independent variables and accuracy as the dependent variable. We found a significant main effect of question type, $F(1,87) = 48.66$, $p < .001$, $\omega^2 = .19$, no main effect of feedback condition, $F(2, 87) = 2.02$, $p = .14$, $\omega^2 = .01$, and no interaction, $F(2, 87) = 1.66$, $p = .20$, $\omega^2 = .01$. Thus, accuracy was significantly lower on leading questions but unaffected by feedback. See Figure 1 below for group/condition means and 95% CIs. See Supplementary Material C1 for an exploratory generalized linear mixed model (GLMM) version of this analysis.

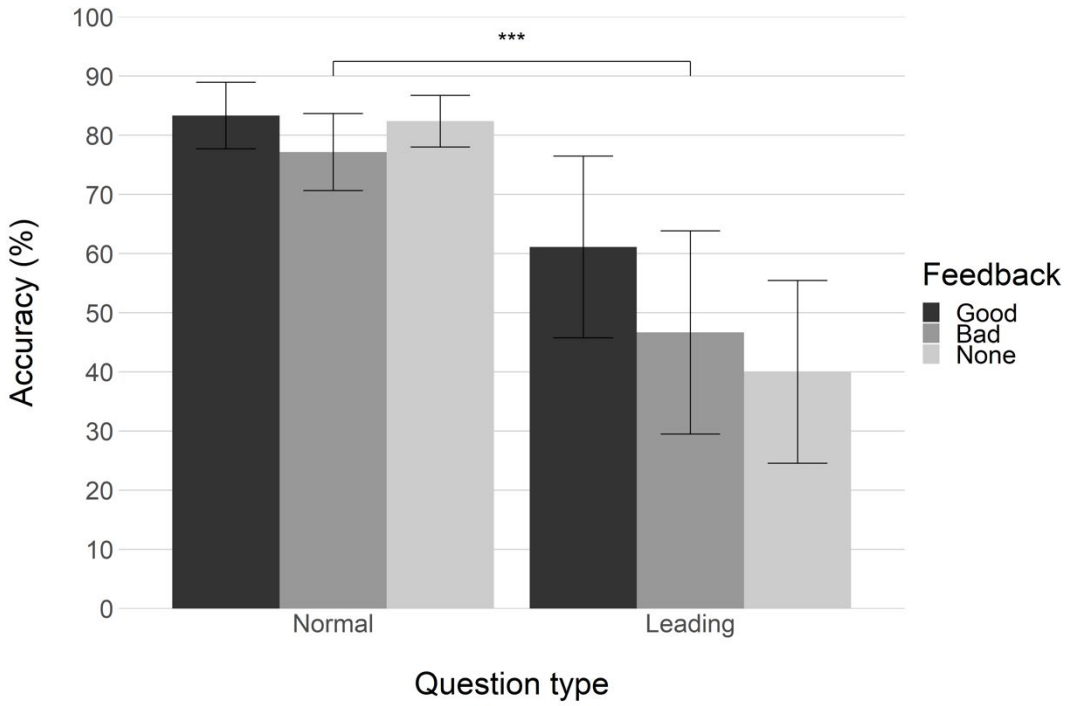

Figure 1. Average accuracy by feedback condition and question type. Error bars represent 95% CIs on each cell mean. ***$p < .001$.

Next, we conducted a similar mixed ANOVA with confidence as the dependent variable. We observed a main effect of feedback condition, $F(2, 87) = 9.14$, $p < .001$, $\omega^2 = .07$, a main effect of question type, $F(1, 87) = 152.31$, $p < .001$, $\omega^2 = .36$, but no interaction, $F(2, 87) = .04$, $p = .96$, $\omega^2 = 0$. Follow-up Bonferroni-corrected *t*-tests[1] comparing the three feedback conditions revealed a significant difference between the good and bad feedback conditions, $t(58) = 4.38$, $p < .001$, Cohen's $d = 1.12$ (95% CI [.57, 1.69]), but no other differences (all ts < 2.39, ps > .02). As with accuracy, confidence was significantly lower on leading questions than on normal questions. Figure 2 below depicts group/condition means and 95% CIs. See Supplementary Material C2 for an exploratory linear mixed model (LMM) version of this analysis, (also including accuracy as a predictor).

[1] Per-test $\alpha = .017$.

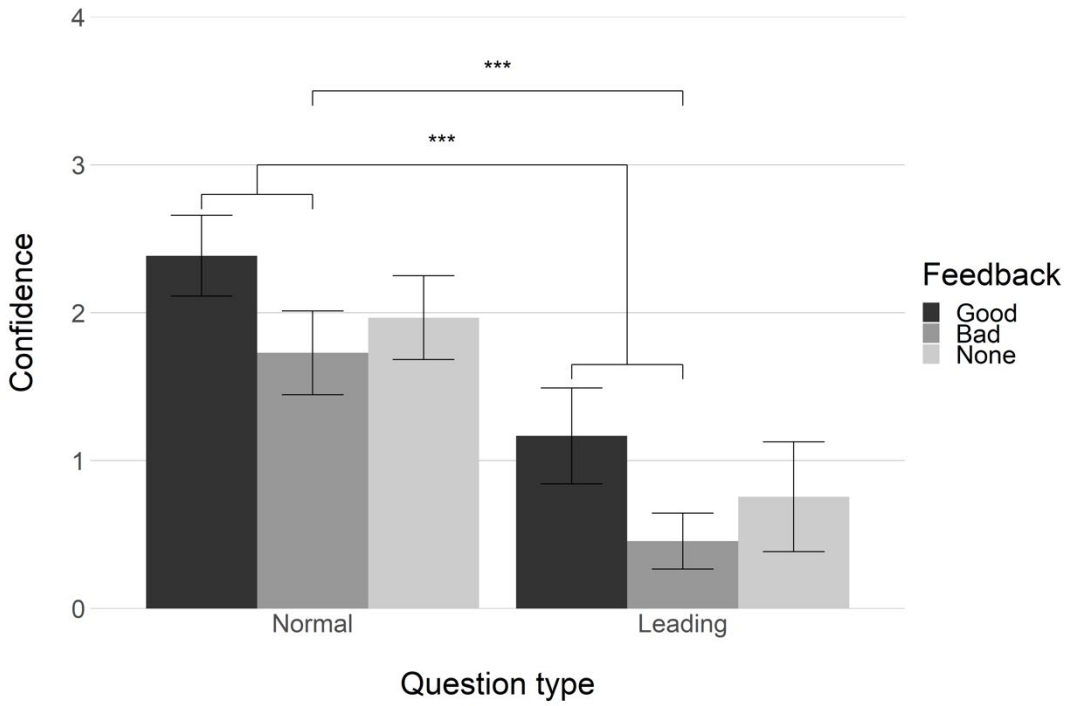

Figure 2. Average confidence by feedback condition and question type. Error bars represent 95% CIs on each cell mean. ***$p < .001$.

Discussion

In this experiment, confidence on both normal and leading questions was affected by feedback—those who received good score feedback were more confident than those who received bad score feedback. However, unlike the results of Leippe and colleagues (2009), there were no differences between our experimental conditions (good/bad score feedback) and our control condition (no feedback). In their experiment, participants who received good score feedback were more confident than those who received no feedback, and participants who received bad score feedback reported lower confidence than those who received no feedback. The nature of the feedback in our experiment may explain why we did not observe the same differences. While Leippe and colleagues (2009) gave participants feedback related to the testimonial accuracy of the event they actually witnessed, we gave participants feedback about their general testimonial accuracy. It is possible that the influence of feedback in our experiment was not as strong as that in Leippe and colleagues' (2009). Despite this, feedback relevant to self-credibility cues seemed to affect confidence judgments, even if the feedback was not directly related to the witnessed event.

In addition, the results of Experiment 1 did not support the predictions of

SCIF—namely, that there would be a significant feedback effect on leading questions only. On the contrary, we found a significant main effect of feedback on leading questions *and* normal questions, where participants should have strong memory traces. Though it is possible that the normal questions were difficult to answer (i.e., memory traces would not be strong), accuracy on normal questions was approximately 80%, suggesting that memory traces were in fact strong. These results imply that participants took external cues (investigator feedback) into account when making confidence judgments even though their internal cues were strong.

Though feedback affected confidence, there were no effects of feedback on accuracy for either question type (normal and leading). These results are consistent with the Cue-Belief model: self-credibility cues influence the assessment of the subjective likelihood that a memory is accurate but not the accuracy of the memory itself. In their experiments, Leippe and colleagues (2006) found that participants who received positive feedback were more accurate in their lineup identification than those who didn't receive any feedback (though they found no evidence for a difference in accuracy between the negative feedback and no feedback conditions). One possible explanation for our lack of accuracy findings was the nature of our feedback manipulation—unlike Leippe and colleagues (2006; 2009), our feedback was not directly related to the event, which may have resulted in a weaker effect than that observed in the original study.

In sum, we found that pre-testimony feedback aimed at manipulating self-credibility cues can influence confidence judgments about a witnessed event. However, such feedback does not seem to affect testimonial accuracy. In Experiment 2, we extended the findings of Experiment 1 by testing the effects of feedback given *after* an initial testimony and *before* a second testimony. Although the influence of feedback following testimony has been studied by many researchers, feedback in these experiments is typically directly related to the witnessed event. Our feedback manipulation allowed us to examine the effects of non-event-related feedback on confidence and accuracy.

Experiment 2

In Experiment 2, our focus was on whether post-testimony self-credibility feedback affects eyewitness accuracy and confidence. In contrast to prior studies, we used general feedback about participants' memory and cognition, feedback that did not directly relate to the witnessed event. The effects of feedback following eyewitness testimony are well-documented, with confirmatory and contradictory feedback generally increasing and decreasing confidence respectively (Wells & Bradfield, 1998;

McGroarty & Baxter, 2007).

However, despite many post-identification feedback studies, almost all the feedback employed is related to extrinsic or intrinsic cues (as opposed to self-credibility cues). Additionally, feedback has typically been related to the actual event witnessed by the participants. Little is known about whether self-credibility cues are affected by feedback that is not directly related to the witnessed event. Using the feedback manipulation from Experiment 1, we examined whether post-testimony feedback affects eyewitness accuracy and confidence. Based on the Cue-Belief model and the results of previous post-identification feedback studies, we predicted that participants in the good score feedback condition have increased confidence in their testimony after receiving feedback, participants in the bad score feedback condition would have decreased confidence, and participants in the no feedback condition would remain roughly the same. As with Experiment 1, we examined whether feedback interacts with question type, as SCIF would predict. Additionally, and in accordance with the Cue-Belief model, we did not predict any effects of feedback on accuracy.

Method

Participants and Design. Participants in Experiment 2 included 90 Japanese students from the University of Tokyo (29 males, 61 females; Age M = 19.43, SD = .87). Our design was a 3 (Feedback condition: good score, bad score, no feedback) x 2 (Question type: Normal, Leading) x 2 (Questionnaire: Pre-feedback, Post-feedback) mixed design, with Feedback condition as a between-subjects IV, Question type and Questionnaire as within-subjects IVs, and accuracy and confidence as separate DVs. Thirty participants were randomly assigned to each of the three feedback conditions. All participants received compensation worth 500 yen.

Material and Procedure. Experiment 2's procedure was almost identical to Experiment 1, with the main difference being the timing of the feedback and the addition of a second post-feedback mock crime video questionnaire. We added a delay task after the first (pre-feedback) mock crime video questionnaire. As with Experiment 1, the experimenter began by explaining that the purpose of the experiment was to examine the validity of a (fictitious) assessment scale predicting eyewitness testimonial accuracy. Participants completed the fictitious assessment scale and then watched the mock crime video. Thereafter, they completed the first video questionnaire. After completing the first video questionnaire, participants worked on a number crossword as a delay task for 10 minutes. After the delay task, participants received their feedback

(good score, bad score, or no feedback). The experimenter then asked participants to complete the second video questionnaire, which was the same as the first one. Before administering the second questionnaire, the experimenter said: "Your report on the previous questionnaire included some wrong answers. I'd like you to try again." The materials used in Experiment 2 were also the same as Experiment 1, except for the questionnaire confidence scale. Because we were able to observe confidence effects on a relatively limited 5-point scale, we changed the 5-point confidence scale to an 11-point scale from 0% to 100% instead of a 5-point confidence scale. We did so to increase our ability to detect more subtle differences (e.g., due to potential anchoring on an initial response) in confidence effects. Participants were informed that 0% meant "Not at all confident" and that 100% meant "Completely confident". Finally, participants were debriefed, told about the study purpose, and probed for suspicion about the fictitious assessment scale. As with Experiment 1, although we cannot fully discount the possibility that some of the participants were not sure if we were specifically asking about suspicion, none of the participants reported any suspicions about the fictitious assessment scale. See Supplementary Material B for scale reliabilities.

Results

Accuracy.

To examine the effects of feedback condition, question type and feedback timing on eyewitness testimonial accuracy, we conducted a three-way mixed ANOVA with feedback condition (good score, bad score, no feedback; between-subjects), question type (normal, leading; within-subjects), and questionnaire (pre-feedback, post-feedback; within-subjects) as the independent variables and accuracy as the dependent variable. This analysis revealed a significant main effect of question type, $F(1, 87) = 98.84$, $p < .001$, $\omega^2 = .30$, a significant interaction between feedback condition and questionnaire, $F(2, 87) = 11.66$, $p < .001$, $\omega^2 = .01$, and a significant three-way interaction between feedback condition, question type, and questionnaire, $F(2, 87) = 7.30$, $p = .001$, $\omega^2 = .01$. No other main effects or interactions were significant (all Fs $< .66$). See Supplementary Material D1 for an exploratory GLMM version of this analysis.

To follow up on the three-way interaction observed in the ANOVA, we conducted two separate two-way ANOVAs (one for each question type), with feedback condition and questionnaire as the independent variables and accuracy as the dependent variable. Effects for each ANOVA were evaluated against a Bonferroni-corrected α

= .025.

For normal questions, there were no significant main effects or interactions (all Fs < 2.10, all ps > .12). For leading questions, there was a significant interaction between feedback condition and questionnaire, $F(2, 87) = 10.26$, $p < .001$, $\omega^2 = .03$, but no main effect of either feedback condition or questionnaire (Fs < .56, ps > .57). We followed up on the interaction using three paired-samples t-tests comparing pre-feedback accuracy and post-feedback accuracy (one t-test per feedback condition, each evaluated against Bonferroni-corrected $\alpha = .017$). Good feedback after the first questionnaire improved accuracy on leading questions on the second questionnaire, $t(29) = 4.26$, $p < .001$, $\Delta_{accuracy} = 18.89$ (95% CI [9.83, 27.95]). Leading question accuracy on the second questionnaire was not affected by bad feedback, $t(29) = .87$, $p = .39$, $\Delta_{accuracy} = 5.55$ (95% CI [-.7.55, 18.66]). However, receiving *no* feedback after the first questionnaire actually reduced leading question accuracy on the second questionnaire, $t(29) = 2.84$, $p = .008$, , $\Delta_{accuracy} = 13.33$ (95% CI [3.75, 22.92]). Figure 3 below depicts group/condition means and 95% CIs.

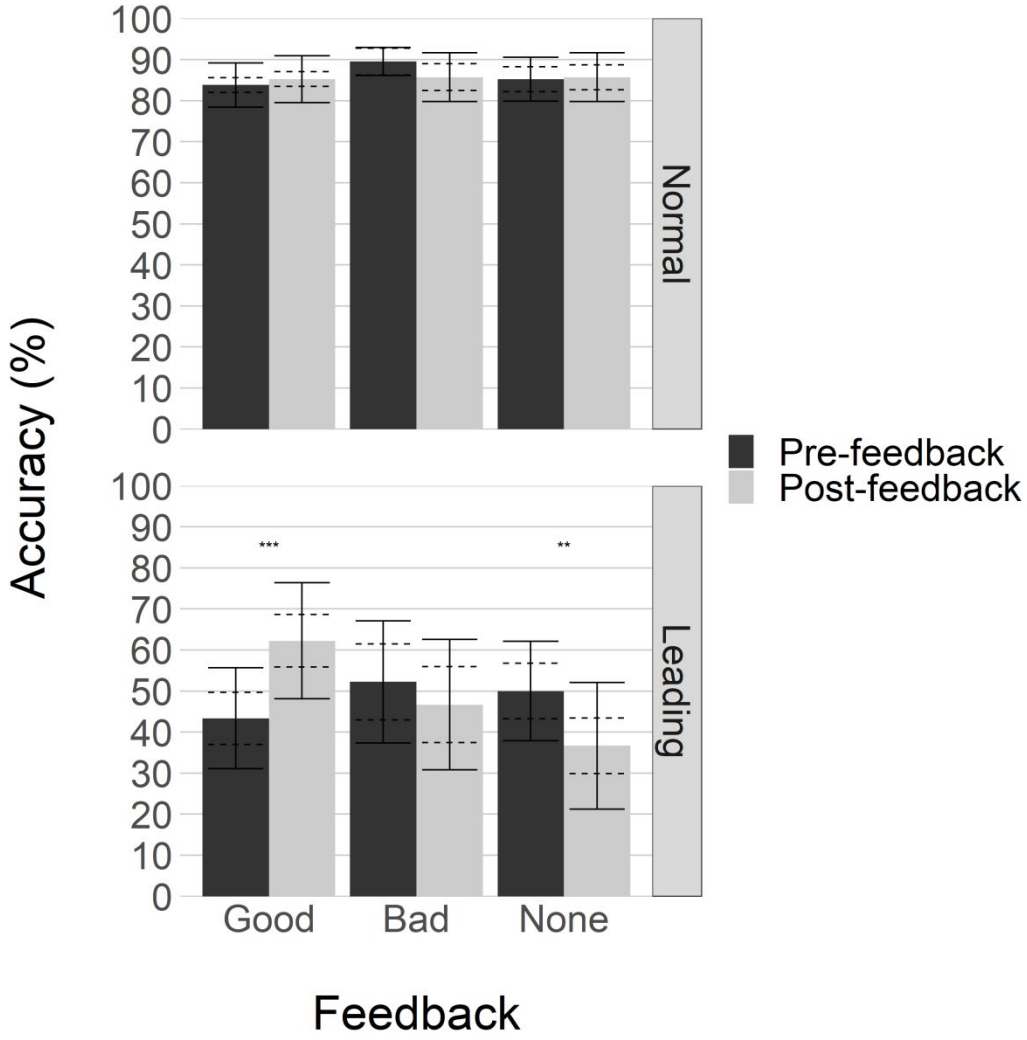

Figure 3. Average accuracy by feedback condition, question type, and questionnaire. Solid error bars represent 95% CIs on individual cell means, while dashed error bars represent within-subjects 95% CIs based on each pre-post comparison (calculated as per Loftus & Masson, 1994). ***$p < .001$, **$p < .01$.

Finally, to address the possibility that there were *pre*-feedback differences in the conditions (i.e., due to experimenter influence), we conducted two final one-way ANOVAs (one for each question type) with feedback condition as the independent variable and accuracy on the first questionnaire as the dependent variable (evaluated against a Bonferroni-corrected $\alpha = .025$). Normal and leading question accuracy on the first questionnaire did not significantly differ by feedback condition (Fs < 1.61, ps > .20). These results failed to provide evidence for pre-feedback differences across

conditions.

Confidence

Using a mixed ANOVA paralleling the one conducted for accuracy, we examined the potential effects of feedback condition, question type and feedback timing on eyewitness testimonial confidence. We found a main effect of question type, $F(1, 87) = 214.93$, $p < .001$, $\omega^2 = .28$, an interaction between feedback condition and questionnaire, $F(2, 87) = 10.81$, $p < .001$, $\omega^2 = .01$, and a three-way interaction between feedback condition, question type and questionnaire, $F(2, 87) = 9.68$, $p = .001$, $\omega^2 = .01$. See Supplementary Material D2 for an exploratory LMM version of this analysis (also including accuracy as a predictor). Like the follow-up analyses for accuracy, we followed this three-way interaction with two separate two-way ANOVAs (one for each question type, effects evaluated against Bonferroni-corrected $\alpha = .025$).

For normal questions, there was a significant main effect of questionnaire such that confidence was significantly but not meaningfully lower on the second questionnaire, $F(1, 87) = 7.45$, $p = .008$, $\omega^2 = .001$[2]. There were no other significant main or interaction effects (Fs < .94, ps > .39). For leading questions however, there was a significant interaction between feedback condition and questionnaire, $F(2, 87) = 11.83$, $p < .001$, $\omega^2 = 0.03$. Separate paired-samples t-tests (one t-test per feedback condition, each evaluated against Bonferroni-corrected $\alpha = .017$) revealed that confidence on leading questions increased after good feedback, $t(29) = 4.01$, $p < .001$, $\Delta_{confidence} = 11.11$ (95% CI [5.45, 16.78]), decreased after bad feedback, $t(29) = 2.65$, $p = .01$, $\Delta_{confidence} = 8.67$ (95% CI [1.97, 15.36]) and remained the same with no feedback, $t(29) = .66$, $p = .51$, $\Delta_{confidence} = 1.78$ (95% CI [-3.70, 7.25]). Figure 4 below depicts group/condition means and 95% CIs.

[2] Due to the miniscule effect size here, we are cautious about concluding any pre/post-feedback differences in normal question accuracy. However, examining the within-subjects contrasts, it appears that the effect was small, but somewhat reliable due to the within-subjects nature of the variable and relatively little within-subjects variability in pre- and post-feedback scores.

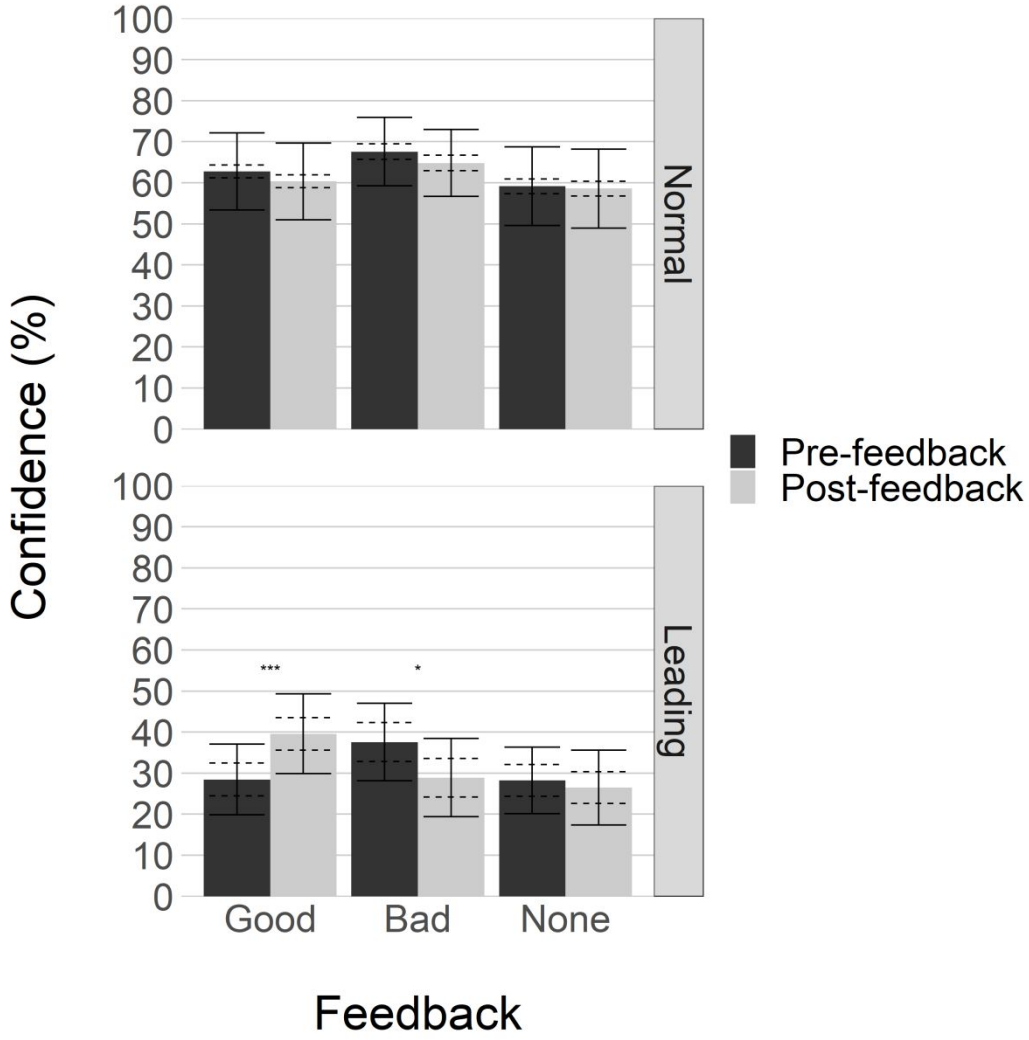

Figure 4. Average confidence by feedback condition, question type, and questionnaire. Solid error bars represent 95% CIs on individual cell means, while dashed error bars represent within-subjects 95% CIs based on each pre-post comparison (calculated as per Loftus & Masson, 1994). ***$p < .001$, * $p < .05$.

It appeared that there may have been slight differences in pre-feedback confidence across conditions, particularly for leading questions. To investigate any potential *pre*-feedback differences in the conditions, we conducted two final one-way ANOVAs (one for each question type) with feedback condition as the independent variable and confidence on the first questionnaire as the dependent variable (evaluated against a Bonferroni-corrected $\alpha = .025$). Normal and leading question confidence on the first questionnaire did not significantly differ by feedback condition (Fs < 1.56, ps

> .21), suggesting no substantial (or at least statistically reliable) pre-feedback differences across conditions.

Response changes. We also performed a chi-square test for homogeneity to examine whether feedback affected the pattern of response changes (e.g., correct-to-correct vs. correct-to-false vs. false-to-correct vs. false-to-false) on the second testimony. We conducted each analysis individually by question type (normal vs. leading). There were no significant deviation of ratios on the normal questions, $\chi^2(6)= 8.94$, $p = .18$, $V = .08$. However, the deviation for the leading questions was significant, $\chi^2(6) = 34.45$, $p < .01$, $V = .25$ (see Tables 2 & 3). Follow-up residual analyses revealed that the proportion of correct-to correct responses was significantly smaller than predicted and the proportion of false-to-correct responses was significantly larger than predicted in the good score condition ($p < .01$). In the no feedback condition, the proportion of correct-to-false responses was significantly larger than predicted ($p < .05$) and the proportion of false-to-correct responses was significantly smaller than predicted ($p < .01$).

Discussion

Effects on accuracy. Though we found that accuracy on both the first and second questionnaires was similar across feedback conditions for both normal and leading questions, receiving good feedback after the first questionnaire increased accuracy on leading questions on the second questionnaire. Conversely, receiving *no* feedback decreased accuracy on leading questions on the second questionnaire.

In addition, the patterns of response changes on the leading questions varied as a function of the kind of feedback received. Participants who received good score feedback improved their testimonial accuracy for leading questions. However, without feedback, accuracy on leading questions was lowered. This is likely because in the no feedback condition, the experimenter stated that the participant's initial testimony included some wrong answers. Presumably, most participants had difficulty correctly answering the leading questions with "None/That feature did not appear in the video." due to their weak memory traces, even if they thought it to be the correct answer. Nevertheless, participants who changed their responses in the good score feedback condition seemed to answer in the negative with more confidence. They may have judged their initial correct response ("None/That feature did not appear in the video") as correct as a result of receiving good score feedback regarding self-credibility cues before the second testimony. In contrast, participants who received no feedback likely

had little confidence in their first testimony and even less in their second testimony as a result of experimenter instructions. Meanwhile, the participants who received bad score feedback were motivated to change their response whether their testimony was correct or not owing to the bad score feedback, which implied that their testimony was not reliable. As a result, there was no significant deviation of response changes in the bad score condition.

Effects on confidence. Feedback had differential effects on confidence for normal and leading questions; for normal questions, feedback did not appear to affect confidence across questionnaires. However, for leading questions, receiving good score feedback after the first questionnaire increased confidence on the second questionnaire, receiving bad feedback decreased confidence, with no feedback leaving confidence unchanged.

Thus, feedback about self-credibility cues—feedback *not* directly related to the witnessed event—had a significant effect on eyewitness confidence for leading questions. The fact that feedback did not affect confidence on normal questions on the second questionnaire is consistent with SCIF: according to SCIF, we assess our confidence based on internal cues. When our internal cues are weak, we look for external cues and take into account the ones we think are reliable. In terms of SCIF, the results of Experiment 2 can be explained thusly: participants seemed to employ only internal cues in assessing their confidence on normal questions, where they would have strong memory traces that would not be influenced by the experimenter's feedback. In contrast, they may have made use of both internal and external cues when judging their confidence on leading questions, where they had no memory traces. On leading questions (with weak or nonexistent memory traces), good feedback served as an external self-credibility cue to increase confidence, while bad feedback served as an external self-credibility cue to decrease confidence.

Thus, our confidence judgment processes may vary as a function of the strength of our memory traces. However, there are at least two other reasons for the absence of feedback effects on normal questions. First, it is possible that the instructions, which implied that there were wrong answers on the first questionnaire, introduced bias that encouraged participants to think that their answers on leading questions were wrong. In fact, a question type (normal vs. leading) × response change (presence vs. absence) chi-square test for homogeneity revealed that the proportion of response changes on leading questions was significantly higher than that of normal questions ($\chi^2(1) = 23.29$, $p < 0.01$, $\varphi = 0.16$). Second, there may be a ceiling effect for

confidence ratings on the normal questions. In the first questionnaire, accuracy on normal questions was extremely high in all conditions (M = 86%, 95% CI [85%, 90%]), whereas accuracy on leading questions was near chance level (M = 47%, 95% CI [41%, 56%]). Though overall confidence ratings on normal questions were not at scale ceiling, it is possible that even highly confident participants were reluctant to report very high confidence levels (i.e., a potential task ceiling). The discrepancy between average accuracy on normal questions and average confidence on normal questions suggests that this is a possibility. One explanation for this potential task ceiling comes from research on cross-cultural differences in confidence in decision-making showing that people in East Asian cultures (including Japan, where this study was conducted) may be less confident in their abilities than Western, individualistic cultures (Mann et al., 1998).

Despite these possibilities, the lower confidence ratings for leading questions permitted more room for changes. In sum, we suggest that good score feedback increased confidence on leading questions but did not increase confidence on normal questions because confidence was already high.

General Discussion

Across two experiments, feedback related to self-credibility cues affected eyewitnesses' confidence judgments. Therefore, it is likely that self-credibility cues play a role in the production of confidence judgments, as the Cue-Belief Model suggests. Furthermore, we found that feedback that was not directly related to the witnessed event had significant effects on eyewitness confidence in both experiments, suggesting that feedback can affect confidence even if the feedback is not directly related to the witnessed event. However, unlike the results of Leippe and colleagues (2009), Wells and Bradfield (1998), and McGroarty and Baxter (2007), we did not find all of the exact expected differences between our conditions (i.e., good score confidence > no feedback confidence and no feedback confidence > bad score confidence). This suggests that feedback effects may be moderated by the degree to which feedback is related to the witnessed event (i.e., less relatedness may = weaker feedback effects).

Additionally, the results of Experiment 1 and 2 were not identical in terms of feedback effects and feedback timing. In Experiment 1, where the experimenter gave feedback before participants testified, participants who received good score feedback reported higher confidence than those who received bad score feedback on both normal and leading questions. On the other hand, in Experiment 2, where feedback was provided after participants testified, feedback only affected confidence on leading questions.

Our question type results can be explained using either the Cue-Belief Model or SCIF. The results of Experiment 1 suggest that we use self-credibility cues when judging our confidence—a result consistent with the Cue-Belief Model. Nevertheless, we could not explain the results of Experiment 2 in terms of the Cue-Belief Model, since the Cue-Belief Model does not explicitly predict question type differences (i.e., normal vs. leading) in self-credibility cues and confidence judgments. However, the results of Experiment 2 can be explained by the SCIF, which posits that we use external cues only when our internal cues are weak (e.g., when we are given leading questions). Though the SCIF readily explains the results of Experiment 2, the same is not true of Experiment 1, where there was a significant feedback effect on normal questions (where memory traces should be strong).

The differential patterns of results, as examined in the context of the Cue-Belief Model and the SCIF, suggest several theoretical implications for confidence judgment processes. First, our confidence assessments are subjective ratings of our memories, and we employ various cues when judging our confidence (as per the Cue-Belief Model). However, our results suggest that accuracy judgments may be less affected by self-credibility cues. Perhaps accuracy is more reliant on internal intrinsic cues (i.e., the familiarity of the memory, which may be less malleable to self-credibility or extrinsic cues). This is consistent with work showing no effects of feedback on accuracy (Steblay, 2014). It is also possible that the susceptibility of accuracy to non-intrinsic cues depends on the directness of the feedback—Leippe and colleagues (2009) found effects on accuracy with feedback directly related to the event in question. Second, our confidence judgments are generally based on both external cues and internal cues, regardless of the strength of our internal cues (as per our results). Third, confidence judgments (and to a lesser extent, accuracy) can be manipulated via self-credibility cues relevant to more general memory abilities. Fourth, we might put more emphasis on internal cues than external cues when internal cues are extremely strong (as per the SCIF). In light of these implications (particularly the fourth), we can readily explain why feedback influenced participants' confidence judgements on leading questions in both Experiments. However, an additional explanation is needed in order to interpret the difference between the feedback effect on normal questions in Experiment 1 and the lack thereof in Experiment 2. The SCIF would predict minimal effects on normal questions in both cases because participants should have strong internal cues with high ecphoric similarity for these questions.

One way to explain these discrepant results is the private thought effect (Wells & Bradfield, 1999). Wells and Bradfield (1999) found that the post-identification effects

were mitigated by participants' pre-identification private thoughts about how sure they were that they would identify the right person in the photospread, how much they focused on the culprit's face and so on. In Experiment 2, participants completed the questionnaire about the witnessed event, rated their confidence for each question, and then received feedback before answering the second questionnaire. It is possible that confidence judgments on the first testimony played a similar role to the private thoughts employed in Wells and Bradfield (1999) in modifying post-testimony feedback effects.

If this is the case, why did private thoughts and confidence judgments on the first testimony mitigate post-feedback effects? Presumably, as Wells and Bradfield (1999) suggest, participants in this study enhanced the strength of their internal cues as a result of their initial confidence judgments and resulting private thoughts. To sum up this possibility and the assumptions about our confidence judgments mentioned above: in Experiment 1, we found significant differences on both normal and leading questions because participants considered both their internal cues and the external cues (experimenter's feedback). In Experiment 2, there was only a significant difference on leading questions. This was because participants' internal cues were intensified by the first confident testimony. For the normal questions, where the original internal cues were strong, participants took only their intensified internal cues into account when making their confidence judgments. However, for leading questions where participants had no memory traces, participants' internal cues were not intensified by the first testimony. Therefore, participants utilized external cues (i.e., experimenter feedback) when making their confidence judgments.

We found that feedback related to self-credibility cues only affected accuracy on leading questions in Experiment 2, while feedback significantly influenced testimonial confidence in both Experiments 1 (for both question types) and 2 (for leading questions). Brewer and Wells (2006) referred to the that factors that influence *only* confidence or accuracy (but not both) may explain prior findings of poor CA relationships. Drawing on the wealth of research on the feedback effect (Steblay et al., 2014), feedback is one factor that may affect confidence but not accuracy. However, in addition to our study, there are a few studies that have found effects of feedback on eyewitness accuracy (e.g., Roper & Shewan, 2002). We suggest that although feedback chiefly influences eyewitness confidence, it can also affect eyewitness accuracy—especially if both our accuracy and confidence judgments are influenced by cues relevant to memory strength, as Leippe and colleagues (2009) assumed with the Cue-Belief Model. From this point of view, it is plausible to assume that feedback affects both confidence and accuracy, but has greater effects on confidence. Intuitively, it

strikes us that the strength of a manipulation required for a quantitative shift in confidence (i.e., 50% confident to 40% confident) is much less than the strength required for a qualitative shift in accuracy (i.e., correct-to-incorrect). On this explanation, the lack of accuracy effects in our study could be due to the fact that our manipulation may have been weaker than that of Leippe and colleagues (2009) due to our focus on general rather than specific meta-memory.

Our results show that feedback about witnesses' general memory capabilities could influence witness' confidence on general event-related questions (in contrast with other studies that have examined effects on identifications). In terms of practical applications, in an ideal world, eyewitnesses should *never* be given feedback (explicit or implicit) about their general or specific memory abilities. However, as we have argued previously, there are numerous potential ways in which eyewitnesses might receive "feedback" about their memory abilities (e.g., scrutiny from the public, other witnesses investigators, or legal personnel). For instance, it is estimated that up to 58% of real-world eyewitnesses discuss the witnessed event with a co-witness (Skagerberg & Wright, 2008)—a situation that could certainly occasion the evaluation and questioning of one's own memory abilities in relation to another. Similarly, if an investigator develops rapport with an eyewitness (as per investigation guidelines given by the US Department of Justice, 1999), the investigator must be careful not to mention the eyewitness' ability. However, it is still possible that an investigator may tell an eyewitness that he/she has a good memory after interviewing (or more likely, inadvertently give implicit positive or negative cues). Though cases of suggestive interviewing have certainly been documented (Garrett, 2011), it is impossible to know the real-world prevalence of investigative interviewing practices that might be considered "feedback" about general memory abilities. However, given the attention and effort devoted to preventing suggestive identification procedures and investigator bias in the legal (e.g., New Jersey Division of Criminal Justice, 2012) and research fields (e.g., Steblay et al., 2014), we believe that our research addresses an issue that has practical implications for eyewitness confidence judgments. For one, it encourages investigators and researchers to pay greater attention to potential sources of both explicit *and* implicit feedback, and to the meta-memorial processes eyewitnesses in the real world engage in (i.e., self-evaluation).

Finally, there are several limitations that are worth considering. First, demand characteristics may have played a role in our experiments. For instance, participants might have figured that the experimenter wanted to behave as if they were a "good" or "bad" eyewitness and may have simply gone along with experimenter feedback. The

experimenter was also not blind to feedback condition. However, given the asymmetrical results obtained in Experiment 1 (effects on both question types) and Experiment 2 (effects on only leading questions) it does not seem plausible that demand characteristics fully explain the differential feedback results. A possible future study could mitigate experimenter effects by providing feedback via computer. Second, the leading and non-leading questions were not counterbalanced. For example, the question about the jacket was always a normal question, whereas the question about the necktie was always a leading question. Also, all leading questions were about the color of the non-existent items—a decision that allowed us to keep the leading questions as similar as possible, but one that may potentially limit to generalizability of our results. However, we think it unlikely that the effects we observed are idiosyncratic to leading questions about the color of clothing items. Finally, performance on the normal questions we used was near ceiling. If these questions were too easy, perhaps memories for them were quite strong and thus most resistant to feedback effects. Despite this limitation, we believe that our study provides evidence for some preliminary boundary conditions—feedback may affect leading questions (poor memory) but not normal questions (good memory).

In conclusion, although our study has some limitations, we have demonstrated the possibility that feedback for self-credibility cues can influence eyewitness confidence both before and after testimony. Furthermore, we suggest mechanisms for how we judge our confidence by pointing out the validity and limitations of two prominent theories.

References

Bradfield, A.L., & Wells, G.L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, 24, 581–594.

Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12, 11–30.

Eisenstadt, D., & Leippe, M. R. (2009). Social influences on eyewitness confidence: The social psychology of memory self-certainty. In R. M. Arkin, K. C. Oleson, & P. J. Carroll (Eds.), Handbook of the uncertain self (pp. 36 – 61). New York NY: Psychology Press.

Garrett, B. L. (2011). Convicting the innocent: Where criminal prosecutions go wrong. Cambridge, MA: Harvard University Press.

Gudjonsson, G.H. (1984). A new scale of interrogative suggestibility. *Personality and Individual Differences*, 5, 303-314.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1304–1316.

Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice*, 3, 7– 41.

Lindsay DS, Read JD, Sharma K. 1998. Accuracy and confidence in person identification: the relationship is strong when witnessing conditions vary widely. *Psychological Science*. 9, 215–18.

Loftus, G.R., & Masson, M.E.J. (1994). Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review, 1*(4), 476-490.

Mann, L., Radford, M., Burnett, P., Ford, S., Bond, M., Leung, K., Nakamura, H., Vaughan, G., & Yang, K-S. (1998). Cross-cultural differences in self-reported decision-making style and confidence. *International Journal of Psychology, 33*(5), 325-335.

McGroarty, A., & Baxter, J.S. (2007). Interrogative pressure in simulated forensic interviews: The effects of negative feedback. *The British Journal of Psychology*, 98, 455465.

New Jersey Division of Criminal Justice (2012). Revised Model Eyewitness Identification Procedure Worksheets.

Quinlivan, D. S., Neuschatz, J. S., Douglass, A. B., Wells, G. L., & Wetmore, S. A. (2012). The effect of post-identification feedback, delay, and suspicion on accurate eyewitnesses. *Law and Human Behavior*, 36, 206–214.

Roper, R., & Shewan, D. (2002). Compliance and eyewitness testimony: do eyewitnesses comply with misleading 'expert pressure' during investigative interviewing? *Legal and Criminological Psychology*, 7, 155–163.

Sauer, J. D., Palmer, M. A., & Brewer, N. (2017). Mock-juror evaluations of traditional and ratings-based eyewitness identification evidence. *Law and Human Behavior*. Advance online publication. http://dx.doi.org/10 .1037/lhb0000235

Skagerberg, E.M., & Wright, D.B. (2008). The prevalence of co-witnesses and co-witness discussion in real eyewitnesses. *Psychology, Crime & Law, 14*(6), 513-521.

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior*, *21*, 283–297.

Steblay, N. M., Wells, G. L., & Douglass, A. L. (2014). The eyewitness post-identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law*, 20, 1–18.

US Department of Justice. (1999). Eyewitness Evidence: A Guide for Law Enforcement.

Wells, G. L., & Bradfield, A. L. (1998). Good, you identified the suspect: Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, 83, 360-376.

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitness' recollections: Can the postidentification-feedback effect be moderated? *Psychological Science*, 10, 138‑144.

Wixted, J. T., Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18, 10–65.

Table 1

Mock Crime Video questionnaire

Question #	Question	Answer type
	Normal questions	
1	What color was the thief's jacket?	Open-ended
2	Was the thief's hair short?	Yes/No
3	What color were the thief's trousers?	Open-ended
4	Did the thief have a mustache?	Yes/ No
5	Did the thief wear a wristwatch?	Yes/ No
7	What color was the bag which the thief stole?	Open-ended
9	What color was the thief's hair?	Open-ended
	Leading questions	
6	What color was the thief's necktie? (The thief didn't wear a necktie.)	Open-ended
8	What color were the thief's gloves? (The thief didn't wear gloves.)	Open-ended
10	What color were the thief's glasses? (The thief didn't wear glasses.)	Open-ended

Note. Participants rated their confidence from 0 (*Not at all confident*) to 4 (*Very confident*) on all questions in Experiment 1, and from 0% (*Not at all confident*) to 100% (*Completely confident*) on all questions in Experiment 2. Though we did not pre-test the questions and their selection was somewhat arbitrary, the leading questions used were roughly based on those used in a previous study (Roper & Shewan, 2002), with slight modifications based on cultural differences (e.g., leading details such as tattoos or studs are less culturally common in Japan, and may have been less likely to be effective as leading questions).

Table 2

Response change patterns on normal questions

Response patterns	Good score	Bad score	No feedback
Correct-to-correct	170	176	167
Correct-to-false	6	12	12
False-to-correct	9	4	13
False-to-false	25	18	18

Table 3

Response change patterns on leading questions

Response patterns	Good score	Bad score	No feedback
Correct-to-correct	38	35	30
Correct-to-false	1**	12	15*
False-to-correct	18**	7	3**
False-to-false	33	36	42

Note. *$p < .05$, **$p < .01$

Table 3

Fictitious assessment scale

Question #	Question
	Section 1
1	When I go to a new place for the first time, I will later be able to recognize the people I meet there.
2	When commuting, there is someone I see and recognize often, even though they don't go to my school or workplace.
3	I focus on passers-by
4	When I memorize other people's faces, I compare them to celebrities.
5	I easily notice when someone changes their hairstyle.
6	When I look at others, I pay attention to their clothes or belongings.
	Section 2
1	I focus on things that other people don't.
2	I make decisions based on intuition.
3	I focus on details first.
4	I understand things better with figures or graphics than with words.
5	I am sensitive to changes in my feelings.
6	When I do something, I devote myself to it.
	Section 3
1	When I do something, I will spend a lot of time completing it.
2	When I do something, I often do it with other people.
3	When I think of doing something, I carry it out.
4	My life has a regular schedule and rhythm.
5	I make plans before acting.
6	I listen more than I speak in daily conversations.