January 1999

"I Know I Know It, I Know I Saw It" : The Stability of the Confidence–Accuracy Relationship Across Domains

Brian H. Bornstein
University of Nebraska-Lincoln, bbornstein2@unl.edu

Douglas J. Zickafoose
Louisiana State University

Published in *Journal of Experimental Psychology: Applied, 5*:1 (1999), pp. 76–88; doi: 10.1037/1076-898X.5.1.76

Submitted November 4, 1996; revised June 30, 1998; accepted July 10, 1998.

"I Know I Know It, I Know I Saw It" : The Stability of the Confidence–Accuracy Relationship Across Domains

Brian H. Bornstein, *Department of Psychology, Louisiana State University*
Douglas J. Zickafoose, *Department of Psychology, Louisiana State University*

If the relationship between confidence and accuracy extended across domains, then one could assess performance in a known domain and use it to estimate performance in another domain. The stability of the confidence-accuracy relationship across the domains of eyewitness memory and general knowledge was investigated. The major findings of Experiment 1 were that in both domains participants were overconfident, yet more confident on correct than on incorrect responses, and that the degrees of overconfidence, calibration, and resolution in the 2 domains were positively correlated. Experiment 2 replicated these findings and showed that feedback about overconfidence reduced overall confidence levels but did not improve calibration or resolution. The implications of these findings are discussed in terms of metamemory and individual differences.

Jurors tend to place a great deal of emphasis on witness confidence in determining witness credibility (Cutler, Penrod, & Dexter, 1990; Fox & Walters, 1986; Luus & Wells, 1994b). Previous research, though, has indicated that witness confidence is only a weak (albeit statistically reliable) predictor of accuracy, with participants generally being overconfident (Berger & Herringer, 1991; Sharp, Cutler, & Penrod, 1988; Smith, Kassin, & Ellsworth, 1989; Sporer, Penrod, Read, & Cutler, 1995).

We thank Emily Elliott and Jeff Wilson for serving as confederates in Experiment 1, and Chris Farrell and Sid O'Bryant for helping with stimulus preparation and data collection in Experiment 2. We are also grateful to Gretchen Chapman, Asher Koriat, Morris Goldsmith, and Lillian Emler for helpful comments on the manuscript, and to Tim Buckley for his statistical advice, insightful comments, and help in developing the general knowledge questions. We also acknowledge the assistance of the MCEG/Sterling film production company, which granted permission to use their film *A Prayer for the Dying* in Experiment 2.

Corresponding author: Brian H. Bornstein.

In addition, confidence and accuracy are influenced by different factors (Luus & Wells, 1994a). This presents a problem, in that jurors may be placing too much emphasis on testimony that is not reliable (Lindsay, 1994). What is needed is a better way to predict witness accuracy.

One possible way would be to determine characteristics of witnesses that are predictive of their accuracy. Deffenbacher (1991) reviewed the literature on the effect of various demographic characteristics on eyewitness reliability and concluded that, with the exception of age, they have only a negligible effect. Deffenbacher concluded that personality traits also have little power to predict either face recognition or event recall, although more recent research (e.g., Hosch, 1994; Kassin, Rigby, & Castillo, 1991) has been somewhat more promising in this respect. For example, Hosch found that high self-monitors are better at face recognition than low self-monitors and that elements of cognitive style, such as field independence, may be predictive of eyewitness accuracy as well. However, evidence supporting the effect

of cognitive styles is mixed (Christiaansen, Ochalek, & Sweeney, 1984; Hosch, 1994).

Another possible way to ascertain how well one's accuracy matches up with one's confidence would be to determine a witness's confidence–accuracy (C-A) relationship in another domain. The most common domain, other than eyewitness memory (EM), used for testing the C-A relationship is participants' confidence in their general, factual knowledge (e.g., Koriat & Goldsmith, 1996; Koriat, Lichtenstein, & Fischhoff, 1980; Liberman & Tversky, 1993; Sniezek, Paese, & Switzer, 1990). The most prevalent finding of these studies is that, as in EM, confidence is a weakly reliable predictor of accuracy, with participants generally being overconfident (Lichtenstein, Fischhoff, & Phillips, 1982). Attempts to discover individual differences in the C-A relationship for general knowledge (GK) questions have also been largely unsuccessful (Lichtenstein et al., 1982; Nelson, 1988; Thompson & Mason, 1996).

There are many ways to measure the C-A relationship, but they generally fall under the headings of either "absolute" or "relative" monitoring effectiveness (Koriat & Goldsmith, 1996; Liberman & Tversky, 1993; Nelson, 1996; Yaniv, Yates, & Smith, 1991). Absolute measures refer to the correspondence between a person's subjective confidence and the proportion correct, such as over/underconfidence and *calibration. Over/underconfidence* compares a person's mean confidence rating to that person's overall accuracy. For example, someone who answers 50% of a set of questions correctly but whose mean confidence rating for that set of questions is 80% would be considered overconfident. In the case of *calibration,* [1] a person would be well calibrated if approximately 70% of all confidence judgments of 70% were actually correct. The main difference between calibration and over/underconfidence is that the former uses the mean of the squared deviations, whereas the latter simply uses the mean deviation. As such, the over/underconfidence measure provides the direction of the relationship in addition to the magnitude, as provided by calibration.

Neither of these two measures is able to assess the extent to which confidence distinguishes correct from incorrect answers, which is the hallmark of relative monitoring measures. *Resolution* accomplishes this purpose by correlating a person's subjective confidence with the correctness of each answer. According to Nelson (1984), the best available measure of resolution is the Goodman–Kruskal gamma correlation, γ. Confidence is positively correlated with accuracy if it is greater for correct than for incorrect responses.

Most of the previous research addressing the C-A relationship has been concerned with absolute monitoring effectiveness, particularly the finding of overconfidence. However, as can be seen from the above discussion, absolute monitoring effectiveness is something quite different from relative monitoring effectiveness (Koriat & Goldsmith, 1996). The difference between the two can be illustrated by people who assign the same confidence level to all of their answers, such as 50%. If these people answered half of a set of questions correctly, then they would show good absolute monitoring effectiveness: They are neither over- nor underconfident (mean confidence and overall accuracy both equal 50%), and they are also perfectly calibrated. However, they would exhibit extremely poor relative monitoring effectiveness because the correct and incorrect responses would both have the exact same confidence ratings.

Despite findings of overconfidence in both the eyewitness and GK areas, surprisingly little research has addressed the relationship between the two domains. Perfect and colleagues (Perfect & Hollins, 1996; Perfect, Watson, & Wagstaff, 1993) compared participants' performance on eyewitness and GK questionnaires. They found that participants were equally overconfident in both domains; however, they did not assess the stability of overconfidence across domains within individual participants. Some support for the notion of cross-domain stability comes from a study by West and Stanovich (1997), who found a significantly positive correlation between participants' degrees of overconfidence in their performance on a GK and on a motor skill task.

Along these same lines, Nelson and Narens (1990) termed the ascription of confidence judgments to information that is retrieved from memory—which is what

[1] The Brier score partition for calibration is $1/N \Sigma\, n(r - c)^2$, where N is the total number of probability assessments, n is the number of probabilities for each category, r is the numerical value of the probabilities for each category, and c is the proportion of probabilities for each category that were attached to the correct alternative.

participants in eyewitness studies are typically asked to do—*retrospective metamemory.* They identified systematic processes in how people make such judgments about the contents of their memories. Thus, monitoring effectiveness in the eyewitness domain can be construed as part and parcel of a larger system that is involved in monitoring memory's contents. Overconfidence in such metamemory judgments might be a relatively stable individual characteristic, similar to cognitive styles such as field independence (Hosch, 1994). If there is a relationship between the degree of overconfidence in the EM domain and the other domain that is used, one could see whether a person was generally over- or underconfident and then generalize to the witnessed event.

The present experiments are an attempt to extend research on the C-A relationship by exploring the stability of individuals' absolute and relative monitoring effectiveness across domains. Of special interest is the question of whether individuals who are good monitors in one domain will likewise tend to be good monitors in the other domain. Finally, we seek to extend the findings of cross-domain stability (West & Stanovich, 1997) by examining the effect that feedback in one domain has on performance in the other domain.

Experiment 1

Given that overconfidence has been found for both GK questions and EM, the main purpose of this study was to determine whether individuals would be stable in their absolute monitoring (i.e., calibration and over/underconfidence) and relative monitoring effectiveness (i.e., resolution) across domains. Participants witnessed a naturalistic event in which two confederates made announcements (cf. Christiaansen et al., 1984). They then completed two unrelated questionnaires, one for GK and one for EM.

On the basis of previous research, we predicted that participants would be overconfident in both the GK domain (Koriat et al., 1980; Liberman & Tversky, 1993; Sniezek et al., 1990) and the eyewitness domain (Berger & Herringer, 1991; Perfect et al., 1993; Smith et al., 1989; Sporer et al., 1995). Second, on the basis of research in both domains showing participants generally to be more confident on correct responses than on incorrect responses (Bothwell, Deffenbacher, & Brigham, 1987; Lichtenstein et al., 1982; Smith et al., 1989), we predicted positive gamma correlations for both GK and memory for witnessed details. Third, research that has found consistency in overconfidence across different domains (e.g., West & Stanovich, 1997) led us to predict that participants' absolute monitoring effectiveness would be stable across the two domains. Finally, although some research has failed to find evidence of stability in resolution across items within a single domain (Nelson, 1988; Thompson & Mason, 1996), findings of stable, systematic processes in people's monitoring abilities in general (Nelson & Narens, 1990)—coupled with the role of personality variables in EM (Hosch, 1994)—led us to the somewhat more tentative prediction of a positive correlation across domains for relative monitoring effectiveness.

Method

Participants

Participants were volunteers from an introductory psychology course at Louisiana State University who received extra course credit. Of the 181 participants who completed the GK questionnaire in Phase 1 of the study, 64 did not provide complete data for analysis, leaving 117 participants for the main analyses.[2] These participants' performance on the GK questionnaire in Phase 1 was compared with that of the 64 participants who were dropped or who did not show up for Phase 2; this comparison yielded no significant differences. Although participants were informed that they would only receive credit for participating in both phases, they were not otherwise forewarned of the importance of the second

[2] A total of 14 participants were dropped for providing unusable data, and 50 participants did not attend Phase 2 of the experiment. Although the number of participants from Phase 1 who did not appear for Phase 2 seems high, it is actually better than the department-wide show-up rate (about 55%) for the semester in which this study was conducted. Another possible reason for this attrition rate may be because Phase 1 was conducted in the first class meeting of the semester, and some of the participants may have dropped the class before Phase 2, thus having no incentive for the extra credit they would have received. The relatively high attrition rate is rectified in Experiment 2.

phase. This was done to keep the study as naturalistic as possible, but it may also explain the relatively high attrition rate.

Procedure

The experiment was conducted in two phases: a GK phase followed by an EM phase. In Phase 1, two confederates addressed an introductory psychology class. One confederate was introduced by the instructor and made an announcement. That confederate then introduced the other confederate, who administered the GK questionnaire. The participants were exposed to both confederates for about 25 min, and each confederate spoke for approximately the same amount of time. In Phase 2, at intervals of either 2 ($N = 70$), 5 ($N = 22$), or 7 ($N = 25$) days later, participants were given the EM questionnaire.[3]

Materials

The study included two measures: A GK questionnaire and an EM questionnaire. Both questionnaires consisted of 46 four-alternative forced-choice questions. Each question was followed by a confidence scale that ranged from 25% (the probability of a correct response by guessing) to 100% by intervals of five. The questions represented a range of difficulty from 7% to 75% correct for the GK questionnaire and from 3% to 100% for the EM questionnaire.

The following are examples of the GK and EM questions:

GK: Ambergris comes from a:

A. Cow	B. Sperm Whale
C. Antelope	D. Elephant

EM: The color of the speaker's shirt was:

A. Blue	B. Gray
C. Green	D. Red

The correct answers to the questions concerning the targets' physical appearance were established by a pilot group while viewing the target individuals.

Results

The participants' overall mean percentage correct and mean confidence were computed for each measure. These means are presented in Table 1, which also shows mean confidence levels on the correct and incorrect responses, mean calibration scores (in all analyses, this score refers to the calibration component of the Brier partition; see Lichtenstein & Fischhoff, 1977), and mean Goodman–Kruskal gamma correlations for each measure. One-way analyses of variance failed to find differences on any of the eyewitness measures that were due to delay, Fs(2, 115) < 1.6, ps > .05, so the data were collapsed across delay intervals for further analysis.

GK Questionnaire

Overall means of confidence and accuracy indicated overconfidence on the GK questionnaire, with participants being 16% more confident on average than they were accurate. The mean calibration score was .26 ($SD = .07$). A calibration curve was constructed with confidence levels being collapsed with the next highest level, such that 25% and 30% were combined, 35% and 40% were combined, and so forth. This curve, shown in Figure 1, indicates overconfidence at every level. The gamma correlations ranged from −1.00 to .80, $M = .21$, $SD = .28$, $p < .01$.

EM Questionnaire

Overall means of confidence and accuracy indicated overconfidence on the EM questionnaire as well, with participants being 19% more confident on average than they were accurate. The mean calibration score was .28 ($SD = .07$). Although participants were both more confident and more accurate on the EM questionnaire than on the GK questionnaire, their global overconfidence and calibration scores on both questionnaires were very

[3] For the sake of realism, participants also made two lineup identifications. Because it is not possible to compute within-subject measures of the C-A relationship for the lineup identifications (unless a very large number of lineups are used), the lineup results are not reported.

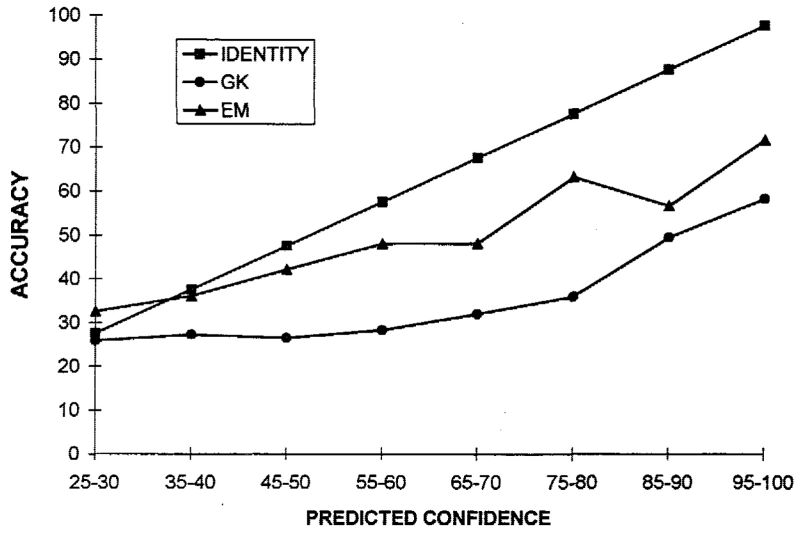

Figure 1. Calibration curves for the General Knowledge (GK) and Eyewitness Memory (EM) questionnaires for Experiment 1

similar. A calibration curve was constructed in the same manner as for the GK scores (see Figure 1). This curve also shows overconfidence, except at the lowest confidence level, for which there were very few responses. The gamma correlations ranged from −.14 to .76, M = .41, SD = .17, $p < .01$.

Correlation Between GK and EM Questionnaire Performance

The overall degree of overconfidence was approximately the same in the two domains: 16% for the GK questionnaire and 19% for the EM questionnaire. Correlations were computed between the two domains for participants' mean confidence, mean accuracy, overconfidence, calibration, and gamma correlation (see Table 2). As predicted, significant positive correlations were found between the GK and EM questionnaires for the absolute monitoring measures (overconfidence, r = .34, $p < .01$; calibration, r = .38, $p < .01$), as well as the relative monitoring measure of gamma (r = .16, $p < .01$). There was also a significant positive correlation for average confidence, r = .33, $p < .01$, but not for average accuracy, r = .09.

Table 1. Mean Accuracy, Confidence, Calibration, and Gamma Correlations for Experiments 1 and 2

	Experiment 1				Experiment 2			
	GK		EM		GK		EM	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Accuracy (%)	31	8	55	8	47	50	50	50
Confidence (%)	47	13	74	11	58	28	60	28
Correct responses (%)	52	15	80	11	68	8	68	8
Incorrect responses (%)	45	14	68	13	50	6	52	6
Calibration score	.26	.07	.28	.07	.25	.05	.26	.08
Gamma correlation	.21	.28	.41	.17	.41	.18	.42	.27

Note. Ns = 117 for Experiment 1 and 96 for Experiment 2. Calibration scores could range from 0 to 1, with lower scores indicating better calibration. GK = General Knowledge questionnaire; EM = Eyewitness Memory questionnaire.

Table 2. Correlations (Pearson's *r*) Between Performance on the General Knowledge (GK) and Eyewitness Memory (EM) Questionnaires

Experiment	Calibration	Accuracy	Confidence	Over/ Underconfidence	Gamma correlation
1	.38**	.09	.33**	.34*	.16*
2	.32**	.16	.49**	.17*	.24

Note. Over/underconfidence is the difference of mean confidence minus mean accuracy.
Ns = 117 for Experiment 1 and 96 for Experiment 2.
*$p < .05$, one-tailed. **$p < .001$, one-tailed.

Discussion

Consistent with previous research, participants were overconfident in answering questions about both impersonal facts (e.g., Lichtenstein et al., 1982) and personally witnessed events (e.g., Smith et al., 1989). They were correct on 31% of their answers to GK questions, yet their mean confidence rating was 47%. Likewise, they were correct on 55% of their answers to questions about the witnessed event, yet their mean confidence rating was 74%. Thus, participants were, on average, 16% and 19% overconfident in the GK and EM domains, respectively.

Although previous research has found similar degrees of overconfidence in GK and EM (Perfect & Hollins, 1996; Perfect et al., 1993), the stability of participants' performance across these domains has not been assessed. The main finding of Experiment 1 is that the same participants who were good monitors in one domain tended to be good monitors in the other domain as well. Specifically, the measures of both relative monitoring (i.e., resolution) and absolute monitoring (i.e., overconfidence and calibration) effectiveness were positively correlated across domains. Of interest, mean confidence was also positively correlated across domains, whereas mean accuracy was not. This finding seems to indicate that participants were merely consistent in their assignment of confidence values, which could account for the positive correlations for the measures of absolute monitoring effectiveness. Although this interpretation can explain the consistency in absolute monitoring effectiveness, the significant positive correlation for resolution between the GK and EM domains indicates that participants for whom differences in confidence reliably predicted differences in accuracy on one task also tended to show good relative monitoring ability on the other task. Thus, participants were not consistent merely in their tendency to use similar confidence values across domains. As a whole, these results suggest that the relationship between an individual's confidence judgments and accuracy, in terms of both absolute and relative monitoring effectiveness, is relatively stable across different tasks (cf. West & Stanovich, 1997).

A practical application of the findings from Experiment 1 would be to present witnesses with feedback, which they would then be able to take into account while testifying. The capacity of feedback to ameliorate the general finding of overconfidence, and potentially to improve witnesses' relative monitoring effectiveness as well, was the primary focus of Experiment 2.

Experiment 2

In Experiment 1, participants were overconfident for both GK and EM questions. Furthermore, both calibration and resolution were found to be moderately correlated across tasks, suggesting a common underlying mechanism controlling performance on both tasks. Experiment 2 was designed to determine whether receiving feedback on GK performance would lead participants to reduce their overconfidence and become more effective memory monitors on an independent task involving EM.

The procedure was very similar to that of Experiment 1, but several methodological changes were made in order to clarify and extend the results. First, the nature of the witnessed event was changed. In Experiment 1, the witnessed event was a live event that had been combined with the GK questionnaire (i.e., part of the witnessed event was the administration of the GK questionnaire). For Experiment 2, the witnessed event was changed to a videotaped clip from a movie. This was done both to

separate the witnessed event from the GK task and to assess the findings' generalizability to a different experimental eyewitness context (cf. Tollestrup, Turtle, & Yuille, 1994).

The next changes from Experiment 1 concerned delay and participant attrition. Because delay did not significantly affect performance on the EM questionnaire in Experiment 1, only a 2-day delay was used for Experiment 2. Additionally, the manner of recruiting participants was altered slightly (see *Participants* below), resulting in a reduction in the attrition rate.

In the final change from Experiment 1, prior to answering questions about the witnessed event, some participants received feedback concerning their performance on the GK questionnaire. Previous studies on the effect of feedback on the C-A relationship have been mixed, depending on the measure used. Some studies have shown that feedback improves resolution but not calibration (Baranski & Petrusic, 1994; Sharp et al., 1988). However, Lichtenstein and Fischhoff (1980) found the opposite result, with feedback improving calibration but not resolution. Subbotin (1996) also found that feedback improved calibration but only for easy items (resolution was not examined in this study). In light of these differences, it is difficult to predict whether calibration, resolution, or both would be improved through feedback. A common finding, however, is that feedback is capable of improving performance (albeit not consistently in all respects).

We used two types of feedback: general feedback, which informed participants of the common findings regarding overconfidence, and specific feedback, which informed them that they themselves had been overconfident in the first phase of the experiment. We made two predictions: (a) that feedback would reduce overconfidence in the eyewitness phase of the experiment, compared with a control condition with no feedback, and (b) that any improvements would be more marked in the specific feedback condition than in the general feedback condition. This second prediction was made because of the heightened relevance of the specific feedback to participants' own behavior.

It is less clear whether feedback (general or specific) about overconfidence would also improve participants' calibration and resolution, as it could lead them to become less confident without any corresponding improvement in how differences in confidence predict differences in accuracy. However, because the feedback might have the overall effect of making participants more thoughtful in using confidence judgments when monitoring their memory performance, we hypothesized that it would improve calibration and resolution as well. As this hypothesis was somewhat tentative, we expected that the feedback about overconfidence would affect participants' calibration and resolution less than their degree of over/underconfidence.

Method

Participants

Participants were volunteers from undergraduate psychology courses at Louisiana State University who signed up to participate in an experiment for extra course credit. Of the 113 participants who completed the GK questionnaire in Phase 1 of the study, 17 did not participate in the EM questionnaire in Phase 2, leaving 96 participants for the main analyses. This attrition rate of 15% is much lower than in Experiment 1.

Procedure

The procedure was similar to Experiment 1 in that it occurred in two phases. In Phase 1, participants, in groups of up to 20, viewed a video clip on a 25-in. monitor and then filled out a GK questionnaire. Participants were instructed to pay close attention to the video because they might be asked about it later. Participants then came back 2 days later for Phase 2. During the delay, the GK questionnaires were scored, and participants were assigned to feedback conditions. In Phase 2, participants were given the EM questionnaire concerning the video, with the first page containing the feedback instructions. Because gains made from feedback have been found to occur following the first feedback session (Lichtenstein & Fischhoff, 1980), only this single instance of feedback was given.

Materials and Design

The GK and EM questionnaires were similar in structure to Experiment 1, but different in content. Specifically, the number of questions on both questionnaires was increased to 50, and the difficulty of the two questionnaires was equated through pilot testing. (The overall accuracy rates for the two questionnaires in Experiment 1 were considerably different—31% for GK, 56% for EM—though performance in both domains was significantly better than chance.) Furthermore, the witnessed event was changed from a live to a videotaped event.

At the beginning of the EM questionnaire was an instruction page that contained one of three feedback conditions: specific, general, or no feedback. In the specific feedback condition, participants were told that their GK questionnaire had been scored in order to provide feedback and that they had been overconfident in their answers. In the general feedback condition, participants were told that some of the participants' GK questionnaires had been scored in order to give them feedback about their performance but that theirs had not been scored. They were then told that most people tended to be overconfident in their answers. The no-feedback (control) instructions simply told participants that the following questions concerned the video they had watched on the first day. Participant triads were matched on their calibration scores on the GK questionnaire, with members of each triad randomly assigned to feedback conditions. This matching process ensured that participants in the different feedback conditions did not differ in their calibration on the GK questionnaire, $F(2, 93) = .331$, $p > .05$.

The witnessed event was a clip about 3½ min long from a popular film. The clip was chosen because it contained a fair amount of dialogue and no violence. The correct answers to questions about the film were determined by unanimous agreement among four raters.

Results

The participants' overall mean accuracy and mean confidence were computed for each task. These means are presented in Table 1, which also shows mean confidence levels on the correct and incorrect responses, mean calibration scores, and mean gamma correlations for each task.

GK Questionnaire

Overall means of confidence and accuracy indicated overconfidence on the GK questionnaire, with participants being 11% more confident on average than they were accurate. The mean calibration score was .25 ($SD = .05$). Mean confidence and accuracy were both higher than in Experiment 1, whereas calibration scores were about the same. A calibration curve was constructed as in Experiment 1 (see Figure 2). This curve indicates overconfidence at every level except the lowest confidence level. The gamma correlations ranged from −.19 to .77, with a mean of .41 ($SD = .18$), $p < .01$.

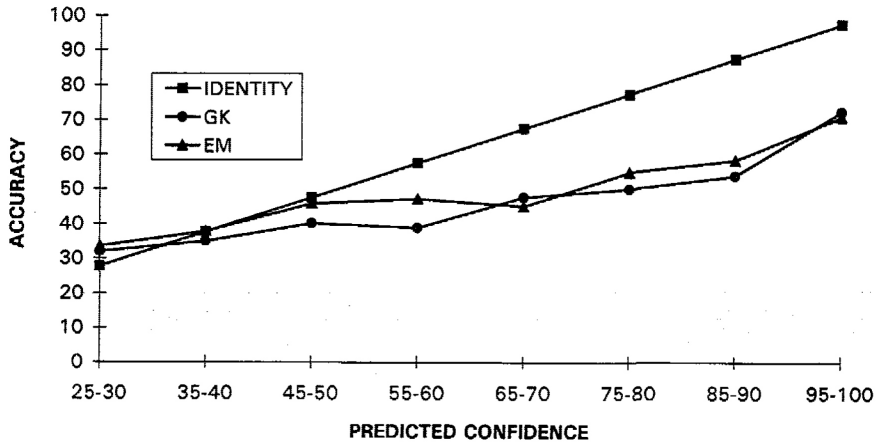

Figure 2. Calibration curves for the General Knowledge (GK) and Eyewitness Memory (EM) questionnaires for Experiment 2

EM Questionnaire

Overall performance. Overall means for confidence and accuracy indicated overconfidence on the EM questionnaire as well, with a mean overconfidence of 10%. The mean calibration score was .26 (*SD* = .08). These figures correspond closely to the GK questionnaire. A calibration curve (see Figure 2) also shows overconfidence, except at the lowest confidence level. The gamma correlations ranged from −.64 to 1.0, with a mean of .42 (*SD* = .27), $p < .01$.

Feedback. Separate analyses of covariance on eyewitness confidence, accuracy, calibration, gamma, and over/underconfidence were conducted, with feedback as a between-subjects factor. Although feedback groups had been equated for calibration on the GK questionnaire, GK accuracy was included as a covariate in all of these analyses to control for any possible variations on this dimension. Type of feedback did not have an effect on eyewitness accuracy, calibration, or gamma (*F*s < 1), but it did have a significant effect on confidence, $F(2, 92) = 8.79$, $p < .01$, and over/underconfidence, $F(2, 92) = 7.56$, $p = .01$. Planned comparisons showed that participants who received either kind of feedback had significantly lower confidence levels than participants who received no feedback (*M*s = 56% vs. 69%), $t(93) = 4.46$, $p < .01$, and were also less overconfident (*M*s = 6% vs. 17%), $t(93) = 2.72$, $p < .01$. There was no significant difference in confidence between the general (*M* = 58%) and specific (*M* = 54%) feedback conditions, $t(93) = 1.09$, $p > .05$; however, participants in the specific feedback condition were marginally less overconfident (*M*s = 2% vs. 10%), $t(93) = 1.92$, $p < .08$.

Correlations Between GK and EM Questionnaire Performance

Because the eyewitness calibration scores and gamma correlations in Experiment 2 did not differ across feedback conditions, all three feedback conditions were collapsed for computing correlations between domains (see Table 2). As in Experiment 1, significant correlations between the GK and EM questionnaires were found for both mean confidence and overconfidence, $r = .49$, $p < .01$, and $r = .17$, $p < .05$, respectively. The correlation between domains for accuracy was marginally significant, $r = .16$, $p < .07$. As in Experiment 1, there were also significant, positive correlations between calibration and gamma on the EM and GK questionnaires, $r = .32$, $p < .01$, and $r = .24$, $p < .01$, respectively.

Discussion

The main result of Experiment 2 is the replication of the significant positive correlations for overconfidence, calibration, and resolution between the GK and EM questionnaires. A second important finding is that, compared to those who did not receive any feedback, participants who were given feedback regarding overconfidence on the GK questionnaire had lower average confidence and overconfidence scores on the EM questionnaire. This reduction occurred whether the feedback indicated that they in particular were overconfident or that people in general were overconfident. However, this reduction in confidence was not accompanied by a corresponding improvement in calibration or resolution. Feedback that was expressed not just in terms of overconfidence but that specifically addressed calibration or resolution, or both, might improve these measures of monitoring effectiveness as well (Lichtenstein & Fischhoff, 1980; Sharp et al., 1988; Subbotin, 1996).

General Discussion

In the present experiments, the relationship between participants' confidence in their memories and the accuracy of those memories was assessed in two different domains: GK and EM. In both experiments, participants' confidence in their responses exceeded their accuracy in both domains, supporting previous research showing that people believe they know more than they actually do about impersonal facts (e.g., Lichtenstein et al., 1982) and personally witnessed events (e.g., Smith et al., 1989; Sporer et al., 1995). More important, the degree to which participants were good monitors was positively correlated in the two domains. Much of this consistency reflected participants' tendency to use similar

confidence ratings across domains; that is, participants' confidence judgments, and not just their degree of overconfidence, were correlated across domains. This consistent use of confidence may account for the stability of absolute monitoring effectiveness (i.e., calibration and over/underconfidence) across domains; however, it cannot explain the stability in participants' resolution scores.

The domains used in the present experiments can be said to draw on distinct memory systems: EM involves episodic memory, which contains experiential memory of events, whereas GK involves semantic memory, which contains abstract knowledge of facts (Tulving, 1983). Tulving has catalogued the extensive differences, as well as the similarities, between these two memory systems. For example, they are proposed to differ in their source and mode of operation, but they are alike in that they both contain information that is propositional in nature and that can be modified as a result of mental activity (Tulving, 1983, chap. 3). The results of the present experiments suggest that another similarity between these two kinds of memory is in people's metaknowledge of the information that is held in episodic and semantic memory. The major finding of the present experiments was that the relationship between participants' confidence and accuracy—in the sense of over/underconfidence, calibration, and resolution—was consistent across domains. Although metamemory has been applied primarily to semantic knowledge (Nelson & Narens, 1990), this finding suggests that it may operate similarly regardless of the type of knowledge that is being monitored. Although some research has failed to find much stability in individuals' metamemory judgments (Nelson, 1988; Thompson & Mason, 1996), both absolute and relative monitoring ability thus appear to be relatively stable characteristics in making confidence judgments across the domains of GK and EM. The stability of metamemory across other tasks also awaits future research.

The results of Experiment 2 suggest that witness overconfidence can be reduced by informing witnesses that people in general (or they themselves) tend to be overconfident. Participants who received such feedback about their performance on the GK questionnaire were significantly less confident in their eyewitness reports than participants who received no feedback. Unfortunately, the feedback did not improve calibration or resolution. In other words, feedback about overconfidence did not affect how well variations in confidence predicted variations in accuracy, despite having the overall effect of reducing participants' confidence.

Although both eyewitnesses and individuals answering questions about impersonal facts vary widely in how well their subjective confidence matches their actual task performance (Smith et al., 1989; Lichtenstein et al., 1982; Luus & Wells, 1994a, 1994b), there are few consistent individual differences in the C-A relationship in either the GK (Koriat et al., 1980; Lichtenstein et al., 1982) or the EM (Deffenbacher, 1991; Hosch, 1994) domain. This dearth of predictors means that it is difficult to determine the degree to which a given individual's confidence is indicative of his or her accuracy. This uncertainty becomes especially problematic when accuracy—that is, the "right" answer—cannot be known conclusively, as is frequently the case in eyewitness situations.

A possible solution to the predictability dilemma that is suggested by the present findings would be to use performance within one domain to predict performance within the other. Specifically, something like a GK questionnaire could be administered to witnesses in an attempt to predict how overconfident they are likely to be in reporting details of the witnessed event. Although jurors are poor judges of eyewitness accuracy (Lindsay, 1994; Wells & Lindsay, 1983), they are nonetheless heavily influenced by eyewitnesses' reported confidence (Cutler et al., 1990; Fox & Walters, 1986; Wells, Ferguson, & Lindsay, 1981). Consequently, they would benefit most—apart from defendants—from learning whether a particular witness tends to be over- or underconfident. The confidence statements of witnesses who were grossly overconfident in responding to GK questions could then be weighed more cautiously than the confidence statements of witnesses for whom confidence and accuracy on GK were more closely related. Such a procedure would capitalize on the finding that

individuals who are overconfident in one domain tend to behave similarly in the other domain. Attending to this fact could therefore correct for witnesses' general tendency to be overconfident. Research showing that jurors are responsive to expert testimony about the unreliability of eyewitness confidence in general (Fox & Walters, 1986) suggests that they would be sensitive to evidence of overconfidence in particular witnesses as well.

This proposal is somewhat limited by the fact that although the correlation between participants' overconfidence scores in the two domains was statistically significant (*r*s = .34 in Experiment 1 and .17 in Experiment 2), it nonetheless means that at most only 12% (according to the *r* of .34) of the variation in eyewitnesses' overconfidence can be explained by the variation in their overconfidence for GK. This is a nontrivial proportion, but it still means that most of the variance is due to other factors, which may include individual differences such as demographic characteristics (Deffenbacher, 1991). Overconfidence for GK questions may be a reliable predictor of overconfidence in eyewitness reports, but it is clearly an imperfect one.

In addition to their practical implications, the present findings have theoretical importance as well. They indicate that there are consistencies in people's metamemory across different judgment domains (cf. Nelson & Narens, 1990). Although situational factors play a considerable role in people's thinking (e.g., Beach & Mitchell, 1978; Payne, Bettman, & Johnson, 1988), the finding that overconfidence, calibration, and resolution are stable in individuals across domains supports theories arguing in favor of cognitive styles (e.g., Wapner & Demick, 1991) and the importance of dispositional factors in how one approaches judgment tasks (Hosch, 1994). Research is called for that addresses in more detail the personality and cognitive factors that are associated with people's monitoring abilities in various metamemory judgment tasks.

References

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, *55*, 412–428.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*, 439–449.

Berger, J., & Herringer, L. (1991). Individual differences in eyewitness recall accuracy. *Journal of Social Psychology*, *131*, 807–813.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, *72*, 691–695.

Christiaansen, R., Ochalek, K., & Sweeney, J. (1984). Individual differences in eyewitness memory and confidence judgments. *Journal of General Psychology*, *110*, 47–52.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior*, *14*, 185–191.

Deffenbacher, K. A. (1991). A maturing of research on the behaviour of eyewitnesses. *Applied Cognitive Psychology*, *5*, 377–402.

Fox, S. G., & Walters, H. A. (1986). The impact of general versus specific expert testimony and eyewitness confidence on mock juror judgment. *Law and Human Behavior*, *10*, 215–228.

Hosch, H. (1994). Individual differences in personality and eyewitness identification. In D. Ross, D. Read, & M. Toglia (Eds.), *Adult eyewitness testimony* (pp. 329–346). Cambridge, England: Cambridge University Press.

Kassin, S. M., Rigby, S., & Castillo, S. R. (1991). The accuracy–confidence correlation in eyewitness testimony: Limits and extensions of the retrospective self-awareness effect. *Journal of Personality and Social Psychology*, *61*, 698–707.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, *114*, 162–173.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149–171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lindsay, R. C. L. (1994). Expectations of eyewitness performance: Jurors' verdicts do not follow from their beliefs. In D. Ross, D. Read, & M. Toglia (Eds.), *Adult eyewitness testimony* (pp. 362–384). Cambridge, England: Cambridge University Press.

Luus, C. A. E., & Wells, G. L. (1994a). Eyewitness identification confidence. In D. Ross, D. Read, & M. Toglia (Eds.), *Adult eyewitness testimony* (pp. 348–361). Cambridge, England: Cambridge University Press.

Luus, C. A. E., & Wells, G. L. (1994b). The malleability of eyewitness confidence and perseverance effects. *Journal of Applied Psychology*, *79*, 714–723.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133.

Nelson, T. O. (1988). Predictive accuracy of the feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (*Vol. 1*, pp. 190–196). New York: Wiley.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item: Comments on Schraw (1995). *Applied Cognitive Psychology*, *10*, 257–260.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–141.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552.

Perfect, T. J., & Hollins, T. S. (1996). Predicting feeling of knowing judgments and postdicting confidence judgments in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *10*, 371–382.

Perfect, T. J., Watson, E., & Wagstaff, G. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology*, *78*, 144–147.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Performance*, *42*, 271–283.

Smith, V., Kassin, S., & Ellsworth, P. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology*, *74*, 356–359.

Sniezek, J. A., Paese, P., & Switzer, F. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, *46*, 264–282.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315–327.

Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, *66*, 268–276.

Thompson, W. B., & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition*, *24*, 226–234.

Tollestrup, P. A., Turtle, J. W., & Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In D. Ross, D. Read, & M. Toglia (Eds.), *Adult eyewitness testimony* (pp. 144–160). Cambridge, England: Cambridge University Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford, England: Oxford University Press.

Wapner, S., & Demick, J. (1991). *Field dependence–independence: Bio–psycho–social factors across the life span*. Hillsdale, NJ: Erlbaum.

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its im-

plications for triers of fact. *Journal of Applied Psychology*, *66*, 688–696.

Wells, G. L., & Lindsay, R. C. L. (1983). How do people infer the accuracy of eyewitness memory? Studies of performance and a metamemory analysis. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 41–55). Chichester, England: Wiley.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*, 387–392.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611–617.