Research Report

Calibration Trumps Confidence as a Basis for Witness Credibility

Elizabeth R. Tenney,[1] Robert J. MacCoun,[2] Barbara A. Spellman,[1] and Reid Hastie[3]

[1]Department of Psychology, University of Virginia; [2]Goldman School of Public Policy and Boalt Hall School of Law, University of California, Berkeley; and [3]Graduate School of Business, University of Chicago

ABSTRACT—*Confident witnesses are deemed more credible than unconfident ones, and accurate witnesses are deemed more credible than inaccurate ones. But are those effects independent? Two experiments show that errors in testimony damage the overall credibility of witnesses who were confident about the erroneous testimony more than that of witnesses who were not confident about it. Furthermore, after making an error, less confident witnesses may appear more credible than more confident ones. Our interpretation of these results is that people make inferences about source calibration when evaluating testimony and other social communication.*

How do you decide whether another person is a reliable source of information? In everyday life, you may rely on the other person's expressions of confidence, history of accurate (and erroneous) statements, and opportunity to be well informed on the subject of the assertion, as well as on your own inferences about the person's motives and character and, in some cases, your beliefs about his or her expertise. Such habits are very general and apply to people's evaluations of everyone, including their acquaintances, friends, and advisors.

Judgments of credibility are especially important in trials. Many behavioral studies show that jurors, or mock jurors, rely on expressed confidence when evaluating eyewitnesses' credibility (Brewer & Burke, 2002; Penrod & Cutler, 1995; Wells, Ferguson, & Lindsay, 1981; Whitley & Greenberg, 1986), despite findings that the correlation between eyewitnesses' confidence and accuracy is weak (Deffenbacher, 1980; Kassin, 1985; Shaw & McClure, 1996). There is even some debate over the "certainty trumps" hypothesis, according to which confidence is the most important determinant of witnesses' credibility (cf. Bradfield & Wells, 2000; Cutler, Penrod, & Stuve, 1988).

A source's past pattern of accuracy is also important in evaluating present credibility. Friendships are lost and business relationships ruined when one party makes too many erroneous statements. In the courtroom, a common strategy to impeach a witness's overall credibility on cross-examination is to demonstrate that the witness has made errors in collateral or central evidentiary assertions (Salhany, 1991; Wellman, 1986). Wigmore (1935, p. 181) cited the aphorism "Falsus in uno, falsus in omnibus" as the rationale for this practice. In the most direct demonstration of this discrediting effect, participants compared the credibility of two witnesses who gave opposing testimony about the party at fault in a routine traffic accident. One witness's nonessential statements about the weather and a peripheral appointment were demonstrated to be inaccurate, and that witness's credibility was substantially reduced (Borckardt, Sprohge, & Nash, 2003). Similarly, a witness's self-contradictory testimony may impeach that witness's credibility (Berman & Cutler, 1996; Hatvany & Strack, 1980), even if the inconsistent testimony is about something trivial (Berman, Narby, & Cutler, 1995; but see Brewer & Burke, 2002, and Lindsay, Lim, Marando, & Cully, 1986, Experiment 3, for nonsignificant effects). Overall, the evidence suggests that errors (both factual errors and self-contradictions) reduce a witness's credibility and the impact of his or her testimony.

We investigated the effects of these two factors, confidence and errors, on witnesses' credibility in mock legal judgments. Past researchers have generally expected additive effects of these factors (see Brewer & Burke, 2002, pp. 354–355). Brewer and Burke manipulated witnesses' confidence via verbal remarks in an audiotaped transcript (e.g., "I am reasonably sure" vs. "I am absolutely sure"). They also introduced four inconsistencies into witnesses' testimony concerning peripheral information about the crime (e.g., direct examination: "the robber was not wearing jewelry"; cross-examination: "the robber was wearing a gold chain"). The witnesses' confidence significantly affected judgments of credibility, but the self-contradictions did not. Most

Address correspondence to Elizabeth R. Tenney, Department of Psychology, University of Virginia, 102 Gilmer Hall, P.O. Box 400400, Charlottesville, VA 22904-4400, e-mail: tenney@virginia.edu.

relevant to our study, there was no interaction between confidence and error. However, the witnesses never made pronouncements of confidence regarding statements later shown to be in error.

In this article, we report two experiments in which eyewitnesses made collateral assertions with high or low confidence and in some cases were shown, by an unimpeachable source, to be incorrect regarding those assertions. Under these conditions, we predicted a nonadditive effect of confidence and error on credibility. In brief, we believe that an error made with high confidence is more damning than a more modest assertion that is equally incorrect because the former evokes inferences about the competency and motives of the source. A high-confidence error implies that the source is either poorly calibrated or dishonestly motivated (see Price & Stone, 2004). A low-confidence error, when it is the only error, implies that the source may be well calibrated about everything. Thus, exhibiting good calibration—being confident when right and unconfident when wrong—should evoke high credibility ratings and be an effective strategy for gaining jurors' support.

In Experiment 1, the collateral evidence concerned eyewitness identification and was closely linked to the central probative claim. In Experiment 2, we pitted two witnesses' testimonies against each other, and the erroneous statement was truly peripheral to the central evidentiary issues.

EXPERIMENT 1

Method

Forty-eight undergraduates (median age = 20; 33 female) completed the 30-min experiment for course credit. Participants read a nine-page written trial summary of a breaking-and-entering case that allegedly occurred on a college campus. Four versions of the case were constructed to create a 2 × 2 between-subjects factorial design with two levels of witness confidence (high vs. low) and two levels of witness error (no error vs. error).

In the materials, an eyewitness testified that he saw the defendant leaving the victim's dormitory room carrying stereo equipment. During cross-examination, the defense attorney asked the eyewitness, "Are you absolutely sure of your testimony?" In the *confident* versions, the witness replied, "Yes, sir, absolutely. I'm certain of it." In the *unconfident* versions, he replied, "No, sir, I'm not certain of it." In the *error* versions, the witness stated that "it was about 7:00," and the defense attorney replied, "You claim that you saw a man leaving [the victim's] dorm room with the stolen goods at about 7:00, but [the victim] has already assured us that he didn't even leave his room until at least 8:15. That seems like a rather serious error to me." In the *no-error* versions, the witness stated that "it was about 8:15," and the defense attorney simply concluded the cross-examination.

All versions of the trial summary contained identical testimony by four other witnesses, as well as instructions regarding presumption of innocence and reasonable doubt. The case was constructed to be sufficiently ambiguous to prevent floor or ceiling effects in verdicts and to ensure that the eyewitness's testimony would play a central role in verdict judgments. After reading the transcript, participants selected a verdict (guilty or not guilty) and rated their confidence in that verdict on a scale from 0 to 10. Participants also rated the witness's believability, trust, and honesty using three 7-point scales, and the average of these ratings was our index of witness credibility (coefficient $\alpha = .71$).

Results and Discussion

Both credibility judgments and verdict preferences show the predicted interaction. A 2 × 2 (Error × Confidence) analysis of variance on credibility ratings revealed a significant crossover interaction, plotted in the top panel of Figure 1, $F(1, 44) = 5.13$, $p_{rep} = .91$,[1] $d = 0.33$. When no error was made, the confident witness was rated more credible ($M = 5.7$, $SD = 1.0$) than the unconfident one ($M = 5.0$, $SD = 1.1$); however, when an error was made, the effects of confidence were reversed, and the unconfident witness was more credible ($M = 5.2$, $SD = 1.0$) than the confident one ($M = 4.6$, $SD = 1.3$).

Participants' verdict preferences were assessed using two measures: a dichotomous (guilty, not guilty) item and a 10-point (0 to 10) confidence-in-verdict scale. We combined these measures by adding 0.5 to each confidence rating[2] and then multiplying the value obtained by −1 in cases of a not-guilty verdict. The resulting 22-point verdict-preference scale ranged from −10.5, indicating complete confidence in a not-guilty verdict, to 10.5, indicating complete confidence in a guilty verdict (cf. MacCoun & Kerr, 1988). A 2×2 analysis of variance on this scale revealed a significant Error × Confidence crossover interaction, $F(1, 44) = 12.43$, $p_{rep} = .99$, $d = 0.51$ (see Fig. 1, lower panel). When no error was made, there was a greater preference for conviction when the witness was confident ($M = 0.7$, $SD = 8.5$) than when he was not ($M = -4.5$, $SD = 7.7$); however, when an error was made, the effects of confidence were reversed, and preference for conviction was greater when the witness was unconfident ($M = 3.2$, $SD = 8.7$) than when he was confident ($M = -5.6$, $SD = 6.3$).

EXPERIMENT 2

Our hypothesis that confidence would moderate the effects of making an error was supported in Experiment 1. The unconfident witness who made an error raised conviction preferences to the same level as the confident witness who did not. In Experiment 2, we wanted to replicate this striking pattern of results with different case materials (civil rather than criminal) and with two witnesses pitted against each other. We also wanted

[1]The p_{rep} statistic indicates the probability of replication, defined as a same-sign result. For further explanation, see Killeen (2005).

[2]There was no no-opinion response option; adding 0.5 to the confidence ratings spread out the scale so that the lowest-confidence guilty and not-guilty responses were 1 point apart.

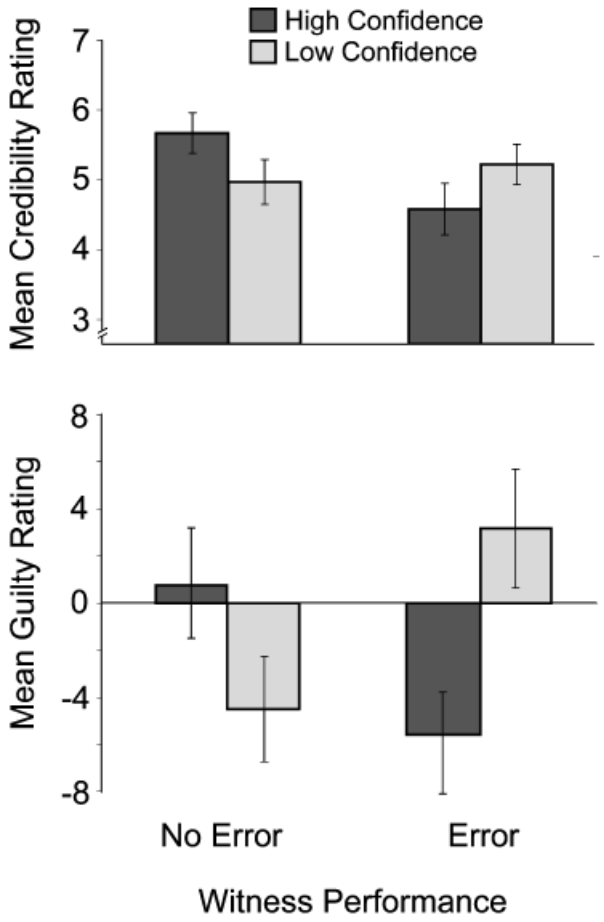

Fig. 1. Results from Experiment 1: mean credibility ratings (upper panel) and mean guilty ratings (lower panel) as a function of the witness's confidence (high or low) and whether or not the witness was shown to have made a collateral error while testifying. The credibility ratings were made on a scale from 1 to 7. The guilty ratings combine verdict preferences with confidence (−10.5 indicates a decision to acquit, made with highest confidence, and +10.5 indicates a decision to convict, made with highest confidence). Because the eyewitness testified against the defendant, lower guilty ratings imply lower credibility of the eyewitness, and higher guilty ratings imply higher credibility of the eyewitness. Error bars indicate standard errors.

to run a study within subjects and across time so that participants could evaluate the credibility of the same witnesses before and after the witnesses made errors. Therefore, participants read testimony from two conflicting witnesses. Each witness described both central and peripheral details of the event; the confident witness was confident about everything, and the unconfident witness was confident regarding the central details but not the peripheral ones. Both were then proven wrong regarding one peripheral detail. As before, we expected that the collateral error would be more damaging to the confident witness than to the unconfident one.

Method

One hundred three undergraduates (median age = 18; 79 female) completed the 20-min experiment for course credit or $5.00. Participants were instructed to act as jurors. They read two fictional depositions in a 2 × 2 within-subjects factorial design with two levels of witness confidence (high vs. low) and two measurements (Time 1, before error, vs. Time 2, after error). The materials were based on Borckardt et al. (2003). At Time 1, participants read two witnesses' contrasting descriptions of a car accident. One witness was not confident about what happened earlier in the day but was confident about how the accident occurred; the other witness was confident about both the details of the day and the accident itself. The witnesses concluded that different vehicles had caused the accident. Participants rated each witness's credibility on a scale from 1, *not credible*, to 6, *credible*, and then chose which witness they believed. The order in which the witnesses were presented was counterbalanced across participants. Participants received no information about the witnesses' sex or race.

At Time 2, participants learned that each witness was correct about the weather conditions on the day of the accident, but that each was wrong about the activity she claimed to have engaged in prior to witnessing the accident. One witness said that she had taken her dog to the veterinarian that day, but the veterinarian's records showed that the appointment had been a week earlier. Similarly, the other witness said that she had gone to a meeting at work about office remodeling, but her boss's records showed that the meeting had been a week later. In light of this new information, participants rerated each witness's credibility and again chose which witness they believed.

Results and Discussion

Both preference judgments (i.e., participants' choice of which witness they believed) and credibility judgments showed the predicted interaction. At Time 1, most participants sided with the confident witness (75.5%), but at Time 2, after the errors were revealed, most sided with the unconfident witness (58.4%); this change was statistically significant, $\chi^2(1, N = 102) = 24.05$, $p_{rep} > .99$, $\phi^2 = .24$.

Results for credibility ratings are shown in Figure 2. At Time 1, the confident witness ($M = 4.5$, $SD = 1.0$) was rated significantly more credible than the unconfident one ($M = 3.7$, $SD = 1.2$), $F(1, 101) = 50.18$, $p_{rep} > .99$, $d = .70$. At Time 2, after participants learned about the errors, the credibility of both witnesses dropped. However, there was a significant interaction between witness confidence and collateral error: Collateral error caused the confident witness to lose more credibility than the unconfident witness, $F(1, 101) = 41.03$, $p_{rep} > .99$, $d = 0.63$. In fact, at Time 2, the unconfident witness was rated as slightly but significantly more credible ($M = 3.0$, $SD = 1.1$) than the confident witness ($M = 2.8$, $SD = 1.0$), $F(1, 101) = 3.90$, $p_{rep} = .88$, $d = 0.20$. There were no significant effects of order of witnesses or sex of participant.[3]

[3] In a subsequent version of the study ($N = 56$), we switched both the activity that each witness claimed to have engaged in on the day of the accident and the witnesses' descriptions of the accident. The pattern of results was replicated.

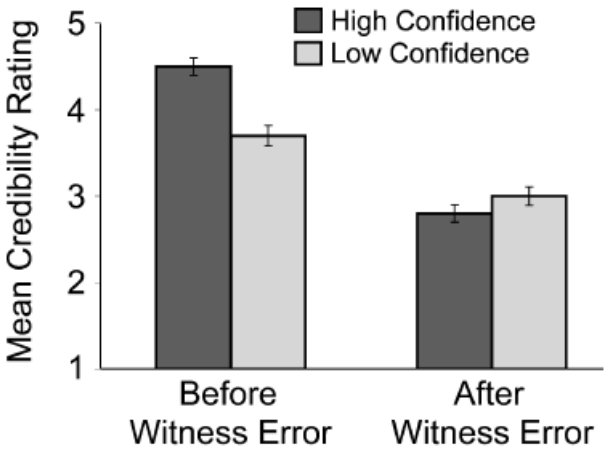

Fig. 2. Results from Experiment 2: mean credibility ratings (1–6) of confident and unconfident witnesses at Time 1 (before each witness made a collateral error) and Time 2 (after each witness made a collateral error). Error bars show standard errors.

Note that, unlike in Experiment 1, the credibility of the unconfident witness was not greater when the witness made an error than when he or she did not. We attribute this difference to a difference in experimental design: In Experiment 2 participants judged the same witnesses before and after learning that each had made an error; in Experiment 1, error was a between-subjects variable.

GENERAL DISCUSSION

People evaluate one another's credibility all the time. Researchers have even proposed that humans have cognitive modules dedicated to the important social tasks of detecting cheaters and liars (Cosmides, 1989). It is not surprising that inferences about credibility are subtle and complicated. The present experiments demonstrate an interaction between a source's accuracy and confidence that affects inferences about the reliability of other statements made by the same source. Observers discount the testimony of a witness who makes a collateral error, even when only some of the witness's assertions are discredited; however, as predicted, a confident witness loses more credibility by making a collateral error than does a witness who expresses uncertainty regarding the erroneous testimony.

There are at least two nonexclusive mechanisms that might produce this pattern. The first involves jurors' inferences about a witness's metacognition: Jurors may use the witness's confidence in erroneous testimony to infer whether the witness is well calibrated. If a witness states that he or she is uncertain about one detail, and then is shown to be in error about it, participants may infer that the witness recognizes the limitations of his or her knowledge. If such a witness then makes a second claim with high confidence, that second assertion would be very believable. In contrast, if a witness expresses confidence about everything, but makes an error, people may infer that the witness is poorly calibrated and may conclude that he or she might be wrong in other high-confidence assertions.

A second mechanism involves jurors' inferences about a witness's motivation. For example, if a witness is obviously motivated to make an assertion, perhaps out of self-interest or prejudice, then the combination of high confidence and error might lead an observer to infer intent to deceive or self-deception on the part of the witness. We believe that this latter account is plausible in many circumstances, but unlikely to be the strongest mechanism in our studies, because our witnesses had no obvious incentive to lie.

More generally, we believe that a comprehensive understanding of how people evaluate the credibility of testimony (or other social assertions) requires an account of the attributions observers make for confidence and errors. A high-confidence error is likely to yield attributions that undermine other statements made by the same source. People giving testimony, advice, or opinions should therefore be careful to express appropriate degrees of confidence in their assertions. Otherwise, the 13th stroke of the clock will cast the other 12 in doubt.

Acknowledgments—We thank Natasha Olinger and Ashley Zapf for help collecting data. This research was supported in part by the National Science Foundation under Grant No. 0437238 to the third author. Portions of this research were presented at the March 2006 American Psychology-Law Society conference in St. Petersburg, FL, and the November 2006 Society for Judgment and Decision Making conference in Houston, TX.

REFERENCES

Berman, G.L., & Cutler, B.L. (1996). Effects of inconsistencies in eyewitness testimony on mock-juror decision making. *Journal of Applied Psychology*, *81*, 170–177.

Berman, G.L., Narby, D.J., & Cutler, B.L. (1995). Effects of inconsistent eyewitness statements on mock-jurors' evaluations of the eyewitness, perceptions of defendant culpability and verdicts. *Law and Human Behavior*, *19*, 79–88.

Borckardt, J.J., Sprohge, E., & Nash, M. (2003). Effects of the inclusion and refutation of peripheral details on eyewitness credibility. *Journal of Applied Social Psychology*, *33*, 2187–2197.

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., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353–364.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *3*, 187–276.

Cutler, B.L., Penrod, S.D., & Stuve, T.E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*, 41–55.

Deffenbacher, K.A. (1980). Eyewitness accuracy and confidence. *Law and Human Behavior*, *4*, 243–259.

Hatvany, N., & Strack, F. (1980). The impact of a discredited key witness. *Journal of Applied Social Psychology*, *10*, 490–509.

Kassin, S. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy-confidence correlation. *Journal of Personality and Social Psychology*, *49*, 878–893.

Killeen, P.R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Lindsay, R.C.L., Lim, R., Marando, L., & Cully, D. (1986). Mock-juror evaluations of eyewitness testimony: A test of metamemory hypotheses. *Journal of Applied Social Psychology*, *16*, 447–459.

MacCoun, R.J., & Kerr, N.L. (1988). Asymmetric influence in mock jury deliberation: Jurors' bias for leniency. *Journal of Personality and Social Psychology*, *54*, 21–33.

Penrod, S.D., & Cutler, B.L. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*, 817–845.

Price, P.C., & Stone, E.R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*, 33–57.

Salhany, R.E. (1991). *Cross examination: The art of the advocate* (2nd ed.). Toronto, Ontario, Canada: Butterworths.

Shaw, J.S., III, & McClure, K.A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*, 629–653.

Wellman, F.L. (1986). *The art of cross-examination: With cross-examinations of important witnesses in some celebrated cases*. New York: Dorset.

Wells, G.L., Ferguson, T.J., & Lindsay, R.C.L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*, 688–696.

Whitley, B.E., Jr., & Greenberg, M.S. (1986). The role of eyewitness confidence in juror perceptions of credibility. *Journal of Applied Social Psychology*, *16*, 387–409.

Wigmore, J.H. (1935). *A student's textbook of the law of evidence*. Mineola, NY: Foundation Press.

(RECEIVED 2/6/06; REVISION ACCEPTED 5/24/06;
FINAL MATERIALS RECEIVED 6/12/06)