Law and Human Behavior
2014, Vol. 38, No. 3, 256–270

0147-7307/14/$12.00 DOI: 10.1037/lhb0000066

Do Confessions Taint Perceptions of Handwriting Evidence? An Empirical Test of the Forensic Confirmation Bias

Jeff Kukucka and Saul M. Kassin
John Jay College of Criminal Justice

Citing classic psychological research and a smattering of recent studies, Kassin, Dror, and Kukucka (2013) proposed the operation of a *forensic confirmation bias*, whereby preexisting expectations guide the evaluation of forensic evidence in a self-verifying manner. In a series of studies, we tested the hypothesis that knowing that a defendant had confessed would taint people's evaluations of handwriting evidence relative to those not so informed. In Study 1, participants who read a case summary in which the defendant had previously confessed were more likely to erroneously conclude that handwriting samples from the defendant and perpetrator were authored by the same person, and were more likely to judge the defendant guilty, compared with those in a no-confession control group. Study 2 replicated and extended these findings using a within-subjects design in which participants rated the same samples both before and after reading a case summary. These findings underscore recent critiques of the forensic sciences as subject to bias, and suggest the value of insulating forensic examiners from contextual information.

Keywords: confirmation bias, context effects, confessions, forensic science

Confirmation biases—that is, the tendency to seek out, interpret, and create new evidence in ways that validate one's preexisting beliefs—are a pervasive psychological phenomenon (Nickerson, 1998). Over the past century, studies using reversible figures (i.e., ambiguous visual stimuli that lend themselves to multiple interpretations; e.g., Boring, 1930; Leeper, 1935), as well as complex stimuli such as blurred photographs (Bruner & Potter, 1964), human faces (Bressan & Dal Martello, 2002), and degraded speech recordings (Lange, Thomas, Dana, & Dawes, 2011), have demonstrated that the perception of a stimulus is often shaped by the observer's expectations about the stimulus. Although some have argued that such biases are enhanced when the observer is motivated (e.g., Balcetis & Dunning, 2006, 2010; Dunning & Balcetis, 2013), others note that confirmation biases are a natural aspect of human cognition that operate outside of conscious awareness (e.g., Klayman & Ha, 1987; Kunda, 1990).

The Forensic Confirmation Bias

Kassin, Dror, and Kukucka (2013) recently reviewed burgeoning evidence that confirmation biases can pervade the judicial system by influencing decision making at both the investigative and adjudicative levels. First, research shows that once a suspect is identified, investigators tend to seek out additional inculpatory evidence and to discount or overlook exculpatory evidence (O'Brien, 2009)—a phenomenon commonly known as "tunnel vision" (Findley & Scott, 2006). For example, investigators perceive more similarity between a suspect and a facial composite when they believe the suspect to be guilty (Charman, Gregory, & Carlucci, 2009), individuals rate detailed statements as less believable when they believe that the source is motivated to lie (Johnson, Bush, & Mitchell, 1998), and interrogators who presume that a suspect is guilty conduct more aggressive interrogations (Kassin, Goldstein, & Savitsky, 2003; Narchet, Meissner, & Russano, 2011) which, in turn, elicit defensive behavior from the suspect that is interpreted as indicative of guilt (Hill, Memon, & McGeorge, 2008; Kassin et al., 2003).

Similar biases guide the subsequent collection and assessment of evidence. There is now literature to suggest that beliefs about a suspect's guilt can impact eyewitness identifications and confidence (Hasel & Kassin, 2009) and evaluations of inconclusive polygraph test results (Elaad, Ginton, & Ben-Shakhar, 1994). At trial, a judge's belief in the defendant's guilt affects how he or she delivers instructions to the jury (Halverson, Hallahan, Hart, & Rosenthal, 1997), and how jurors evaluate otherwise ambiguous evidence. In one study, for example, mock jurors "heard" more incriminating information in degraded audio recordings when led to believe that the speaker was a criminal suspect (Lange et al., 2011).

Bias in Judgments of Forensic Evidence

Although forensic science evidence has been deemed trustworthy and routinely admitted at trial for nearly a century (Mnookin et al., 2011), a growing body of research indicates that even expert judgments of forensic science evidence are subjective and susceptible to bias. In short, by changing the context in which the evidence is presented, forensic scientists form expectations that can influence their perceptions and judgments.

This article was published Online First December 16, 2013.

Jeff Kukucka and Saul M. Kassin, Department of Psychology, John Jay College of Criminal Justice.

Correspondence concerning this article should be addressed to Jeff Kukucka, Department of Psychology, John Jay College of Criminal Justice, New York, NY 10019. E-mail: jkukucka@jjay.cuny.edu

In the earliest empirical study of this phenomenon, Miller (1984) asked 12 individuals who had been trained in the identification of forged signatures to participate in a mock forgery investigation in which they compared a forged check against handwriting samples from one or more suspects. When given no other information about the case, all participants correctly concluded that none of the suspects had authored the forged check. However, when told that a suspect had been identified by two eyewitnesses, four out of six participants incorrectly concluded that this suspect had forged the signature and a fifth deemed the results inconclusive.

Dror and colleagues (Dror & Charlton, 2006; Dror, Charlton, & Peron, 2006) have demonstrated similar effects among latent fingerprint experts. In one study, Dror, Charlton, and Peron (2006) represented five experienced fingerprint examiners with sets of prints that they, unknowingly, had themselves judged as a match several years earlier. When led to believe that the prints were taken from a high-profile misidentification case, thus leading the examiners to expect that the prints would not match, four of the five now concluded that the prints did not match, thereby changing their own prior evaluations of the same evidence. A second study by Dror and Charlton (2006) represented six experienced examiners each with eight sets of prints that they had previously examined. Half of these were presented with biasing case information—either that the suspect had confessed (implying that the prints would match) or that the suspect had produced a verified alibi (implying that the prints would not match). Overall, 17% of judgments made in the biasing information conditions changed over time, and four of the six experts changed at least one of their earlier judgments.

It appears that even DNA testing—widely considered the "gold standard" in forensic evidence (Lynch, 2003; Saks & Koehler, 2005)—is subject to bias under some circumstances. Dror and Hampikian (2011) described an actual gang rape case in Georgia in which one of three assailants accepted a plea bargain to testify against the other two suspects. Under state law, however, the testimony of the admitted rapist was inadmissible without corroborating evidence. Aware of this stipulation, DNA analysts concluded that a complex DNA mixture taken from the victim's body implicated the other two men. To determine if knowledge of this evidentiary rule may have biased these analysts, the authors later presented the same DNA mixture (along with samples from the suspects and victim) to 17 independent analysts and found that only one agreed with the original conclusion. Twelve concluded that the DNA sample did in fact exclude the other suspects; four judged the samples as inconclusive.

The foregoing research underscores concerns raised by the National Academy of Sciences (NAS) in its 2009 critique of the forensic sciences for their lack of certified training, standardization, reliability, and peer-reviewed empirical research and the "potential for bias and error in human observers" (NAS, p. 8). Consistent with this critique, post hoc analyses of DNA exoneration cases have implicated forensic science errors as a common contributing factor in wrongful convictions (Garrett & Neufeld, 2009; Hampikian, West, & Akselrod, 2011; www.innocenceproject.org).

Handwriting Evidence

One discipline of forensic science that has come under heavy criticism is questioned document examination, colloquially known as handwriting identification, which entails the comparison of handwriting samples to determine whether they were authored by the same individual (i.e., a "match"). The existing empirical literature comparing the performance of experts and laypersons on such a task is scant and rather mixed. Although some researchers (e.g., Kam, Fielding, & Conn, 1997; Kam & Lin, 2003) argue that handwriting experts possess identification skills superior to those of laypeople, others (e.g., Risinger, 2007; Risinger & Saks, 1996) have scrutinized the data on which these claims are based and come away with a less optimistic view. For example, Kam, Fielding, and Conn (1997) found that experts and nonexperts were equally proficient at identifying matches but that experts committed fewer false positive errors. Yet Risinger (2007) noted that the experts in this study also produced high false positive (6.5%) and false negative (12.5%) error rates. The NAS (2009) neatly summarizes the ongoing debate, noting that "there may be some value in handwriting analysis" but that its "scientific basis . . . needs to be strengthened" (pp. 166–167).

Controversy notwithstanding, the presentation of handwriting evidence at trial is not uncommon, and questioned document examiners are often proffered as expert witnesses (see Risinger, 2007, for a compendium of relevant post-*Daubert* cases). In many cases, the testimony of handwriting experts has been excluded under *Daubert*, insofar as it does not qualify as "scientific knowledge" (Daubert v. Merrell Dow Pharmaceuticals, 1993), but deemed admissible (at least in part) under either *Kumho* (i.e., *Kumho Tire Co. v. Carmichael*, 1999) or FRE Rule 702 as specialized, nonscientific testimony that would "assist the trier of fact" (e.g., *U.S. v. Hines*, 1999; *U.S. v. Paul*, 1999; *U.S. v. Starzecpyzel*, 1995). Further, many of these cases echo concerns originally raised in *U.S. v. Buck* (1987), where the defense argued not only that jurors are capable of performing the same visual comparisons as the proffered expert, but that it may in fact be preferable for jurors to perform the analysis on their own to protect against "undue prejudice from the mystique attached to 'experts'" (p. 3).

The Current Studies

The literature reviewed above raises three important questions regarding the use of handwriting evidence. First, to what extent are laypeople capable of performing a visual comparison of handwriting samples and drawing accurate conclusions about authorship? Second, consistent with the operation of forensic confirmation biases (Kassin et al., 2013), would people's visual comparisons of handwriting samples be tainted by knowledge of case information that leads them to presume the suspect's guilt or innocence? Third, in light of research suggesting that confirmation biases are heightened by stimulus ambiguity—consistent with the claim that evidentiary judgments vary in terms of their malleability or "elasticity" (Ask, Rebelius, & Granhag, 2008)—would the biasing effect of case information be constrained by the similarity of the handwriting samples that are being compared?

With these questions in mind, we tested the hypothesis that jurors' perceptual judgments of handwriting evidence would change as a function of other evidence—namely, whether the defendant had confessed during a police interrogation. Confessions are a uniquely potent form of evidence (Kassin & Neumann, 1997) and create a strong belief in the defendant's guilt even when

coerced and retracted, even when jurors are instructed to disregard the confession, and even when they believe it did not affect their decision making (Kassin & Sukel, 1997). Research shows that laypeople have only a limited understanding of the risk factors that produce false confessions (Henkel, Coffman, & Dailey, 2008) and generally believe that false confessions are unlikely, even in coercive interrogations (Leo & Liu, 2009).

Thus, we expected that knowledge of a prior confession, even after the defendant has recanted it, claiming it was coerced, would engender a strong belief in the defendant's guilt. Further, we predicted that individuals who are shown two handwriting samples—one from the defendant and one from the perpetrator—would perceive greater similarity between them, and would be more likely to erroneously conclude that they are a "match" (i.e., were written by the same person) when told that the defendant had previously confessed. Finally, by independently varying the similarity of these handwriting samples, we expected that the effect of the confession on evidentiary judgments would be moderated by the similarity of the handwriting samples, such that its impact would be greater when the samples appeared more similar.

Pilot Study

The purpose of our pilot study was threefold. First, we aimed to pretest and identify pairs of handwriting samples that varied in terms of their perceived similarity. This would allow us to later analyze similarity as a potential moderator of bias. Second, we hoped to obtain some sense of laypeople's baseline ability to accurately evaluate pairs of handwriting samples for similarity and shared authorship. Third, we aimed to show that lay judgments of handwriting evidence are "biasable" by showing a discrepancy between participants' willingness to believe that two samples *could have been* authored by the same person and their willingness to believe that those samples *were in fact* authored by the same person.

Method

Participants. A sample of 55 undergraduates completed the pilot study online in exchange for research credit for their introductory psychology course. No demographic information was collected.

Procedure. Participants were shown a total of 15 pairs of handwriting samples, presented sequentially and in a randomized order using a survey Web site. Each pair was presented without any contextual information; participants were simply instructed to compare the handwriting in the two samples, and then to rate their similarity as well as indicate their belief that the two samples were (or were not) authored by the same individual.

Materials. Each handwriting pair consisted of a *target* sample, which was held constant across all 15 trials, and a *comparison* sample, which varied across all 15 trials. The target sample had been authored by a graduate student, and read, "I understand my rights to remain silent and to call a lawyer and I agree to talk at this time."

For each trial, this target sample was compared against one of 15 comparison samples, all of which read, "I have a gun. Keep quiet or I will shoot you. Give me all your cash!" One of these 15 samples was authored by the same individual who had authored the target sample; this was included to examine participants' ability to accurately detect when two samples were truly authored by the same individual (i.e., a "true match"). The remaining 14 comparison samples were "nonmatches," that is, were not authored by the same individual who had authored the target sample. Included in these 14 nonmatches were two comparison samples authored by the same individual on two different occasions; these were included to examine the consistency of participants' judgments of comparison samples provided by the same author. The remaining 12 comparison samples were authored by 12 different individuals.

Dependent measures. For each pair, participants made three judgments. First, they rated the *similarity* of the handwriting in the two samples, using a scale that ranged from 1 (*not at all similar*) to 10 (*very similar*). Second, they indicated whether or not they believed it was possible that the two samples could have been authored by the same individual (*possible match*). Finally, they indicated whether or not they believed the two samples had in fact been authored by the same individual (*actual match*).

Results

Descriptive statistics. Descriptive statistics for all 15 pairs can be found in Table 1. The overall mean similarity rating across sample pairs was 4.05 ($SD = 2.70$), with means for individual pairs ranging from 2.09 to 5.73 and standard deviations ranging from 1.86 to 3.08. Eleven of the 15 pairs were unimodal with a mode of 1 for similarity rating, and 12 of the 15 pairs had a range of 9, the maximum possible range for the scale. The overall percentage of possible match judgments was 55.37%, with individual pairs ranging from 34.55% to 72.73%. Finally, the overall percentage of actual match judgments was 26.91%, with individual pairs ranging from 9.09% to 45.45%.

From the pairs that were authored by different individuals ("nonmatches"), we selected the two low-similarity pairs (Pairs 4 and 6) and two high-similarity pairs (Pairs 5 and 14) to be used in Study 1 on the basis of their mean similarity ratings (for examples of these stimuli, see Appendix). A repeated-measures ANOVA with post hoc Bonferroni analyses confirmed that the two high-similarity pairs did not differ in terms of perceived similarity, and were rated as significantly more similar than the two low-similarity pairs, which also did not differ from each other, $F(3, 162) = 19.32$, $p < .0001$, $\eta_p^2 = .263$.

Judgment accuracy. To analyze differences in *perceived similarity* between the "true match" and other pairs, a repeated-measures ANOVA was performed, which revealed differences in mean similarity ratings among the 15 pairs, $F(14, 756) = 11.55$, $p < .0001$, $\eta_p^2 = .176$. Post hoc Bonferroni analyses indicated that the "true match" was rated as more similar than six of the other 14 pairs, and as equal in similarity to the remaining eight pairs (see Table 1).

Similarly, 72.73% of participants judged the "true match" pair as a *possible match*. A series of one-way chi-square tests, using .7273 as the expected proportion of possible match judgments, indicated that 10 of the other 14 pairs were less likely, and four pairs were equally likely, to be judged as a possible match, compared with the true match pair. Finally, 32.73% of participants correctly judged the true match pair as an *actual match*. A series of one-way chi-square tests, using .3273 as the expected propor-

Table 1
Descriptive Statistics From Pilot Study of 15 Handwriting Sample Pairs

	Similarity ratings						
Pair	*M*	*SD*	*Mdn*	*Mode*	Range	% "Possible Match"	% "Actual Match"
1	2.09_{a}	1.86	1	1	7	34.55	16.36
2	4.18_{be}	2.70	4	1	9	56.36	23.64
3	4.31_{be}	2.68	4	1	9	60.00	23.64
4	3.02_{ab}	2.25	3	1	9	38.18	10.91
5	5.38_{e}	2.47	5	5	9	70.91	43.64
6	3.15_{abc}	2.01	3	1	7	43.64	9.09
7	2.93_{ab}	2.48	2	1	9	40.00	20.00
8*	5.02_{ce}	2.92	6	1	9	65.45	45.45
9	4.33_{be}	2.48	4	3	9	61.82	29.09
10	3.29_{abc}	2.58	2	1	9	40.00	16.36
11	4.22_{be}	2.41	4	1	9	58.18	27.27
12	3.29_{abd}	2.26	3	1	7	56.36	27.27
13^	4.95_{e}	2.66	5	3,4	9	72.73	32.73
14*	5.73_{e}	3.08	6	1	9	72.73	45.45
15	4.89_{de}	2.70	5	1,6	9	58.18	32.73

Note. Similarity ratings were given on a scale from 1 (*not at all similar*) to 10 (*very similar*). Mean similarity ratings not sharing a common subscript differ at $p < .05$.
^ "True match" pair (i.e., target and comparison samples were authored by the same individual). * Comparison samples for these pairs were authored by the same individual.

tion of actual match judgments, indicated that the true match pair was more likely to be judged as an actual match than five, and less likely to be judged as a match than two, of the other 14 pairs. The true match did not differ from the remaining seven pairs in terms of its likelihood of being judged as an actual match.

Judgment consistency. To examine the consistency of participants' judgments over time, we included two comparison samples that had been authored by the same individual and then compared how these samples were rated against the target. A paired-samples *t* test indicated that although similarity ratings for these two pairs (*M*s = 5.02 and 5.73; *SD*s = 2.92 and 3.08, respectively) did not differ, $t(54) = 1.38$, $p = .173$, there was no significant correlation between these ratings, $r(53) = .19$, $p = .155$. Moreover, there was no relationship between these two pairs in terms of possible match judgments, McNemar $\chi^2(1) = 0.89$, $p = .481$, or actual match judgments, McNemar $\chi^2(1) = 0.00$, $p = 1.00$.

Biasability. Separate logistic regression models indicated that similarity ratings predicted both possible match judgments, $\beta = .772$, Wald $\chi^2(1) = 223.60$, $p < .0001$, $OR = 2.17$, 95% CI [1.96, 2.40], and actual match judgments, $\beta = .727$, Wald $\chi^2(1) = 201.39$, $p < .0001$, $OR = 2.07$, 95% CI [1.87, 2.29]. As depicted in Figure 1, the logit equations produced by these models revealed markedly different predictive trends. As expected, there was a discrepancy between participants' willingness to indicate a possible match and their willingness to conclude that the samples are an actual match (e.g., given a similarity rating of 5, the models predict a 78.53% probability of believing that a match is possible but only a 27.67% probability of concluding that the samples are an actual match).

Figure 1. Similarity rating as a predictor of possible and actual match judgments (pilot study).

Study 1

The pilot study produced three main findings. First, lay judgments of handwriting samples tended to be very conservative. Only three out of 15 pairs generated a mean similarity rating above the midpoint of a 10-point scale, the overwhelming modal response to the samples was "not at all similar," and no single pair was judged an actual match by a majority of participants. Second, judgments were highly variable and largely inaccurate, as evidenced by ranges and standard deviations, inconsistencies in evaluating two comparison samples that were authored by the same individual, and a general inability to distinguish the "true match" from the other pairs. Indeed, two of the nonmatching pairs were more likely to be judged as an actual match than the "true match" pair. Finally, even when participants were unwilling to conclude that two samples were an actual match, they were quite willing to believe that a match was possible, suggesting the potential for contextual information to bias evidentiary judgments.

The aim of Study 1 was to present handwriting samples in the context of a criminal trial to test the hypothesis that knowledge of a prior confession would taint individuals' perceptions of handwriting evidence. Participants read a case summary in which the defendant did not confess or confessed and then recanted his confession. Then they evaluated two handwriting samples—one from the known perpetrator, and one from the defendant—that were either high or low in their pilot-tested similarity. Thus, Study 1 employed a 2 (Confession: Present vs. Absent) × 2 (Handwriting Similarity: High vs. Low) between-subjects design.

Method

Participants and design. A snowball sample of 171 participants was obtained via online social networking media. Participants had a mean age of 28.86 ($SD = 8.81$) and were predominantly White (86.55%) and female (67.84%), and held at least a bachelor's degree (84.12%). Each participant was randomly assigned to one of four cells produced by the 2 × 2 between-subjects design. Two participants (1.17%) were later excluded after failing a manipulation check in which they were asked to identify whether the defendant had previously confessed, leaving a final sample of $N = 169$ for all analyses.

Procedure. Participants completed the study using an online survey Web site and were not compensated for their participation. After providing consent as well as basic demographic information, participants were told that they would read a summary of an actual criminal investigation. They then read one of two summaries of a bank robbery case based on *U.S. v. Hines* (1999). In each summary, an armed man gave a handwritten note to a bank teller and escaped with a large sum of money; the note provided a sample of the perpetrator's handwriting. Police then apprehended a suspect (Johanna Hines), who matched the teller's general description, and brought him to the police station where he handwrote and signed a waiver of his *Miranda* rights and was questioned for three hours. Participants in the *confession-present* condition were then told that Hines gave a detailed confession, which he later recanted prior to trial, claiming he was coerced. Those in the *confession-absent* condition were told that Hines maintained his innocence throughout the interview.

All participants were then asked to imagine that they were members of the jury at Mr. Hines' trial and have been presented with two items of evidence: (a) the note that the perpetrator handed to the bank teller, and (b) the *Miranda* waiver written by Hines, the defendant. By random assignment, participants received either one of two high-similarity pairs (*high-similarity* condition) or one of two low-similarity pairs (*low-similarity* condition) that had been identified from the pilot study. Participants were instructed to compare the samples carefully and given an unlimited amount of time to do so. During this time, they rated the similarity of the two samples, indicated whether they believed them to be authored by the same individual, and made a judgment concerning the defendant's guilt or innocence. Lastly, participants completed a comprehension test designed to ensure that they read, understood, and recalled the case summary.

Case summary. We developed a case summary loosely based on *U.S. v. Hines* (1999), an important post-*Kumho* federal case in which District Judge Nancy Gertner ruled that expert testimony by a questioned document examiner was admissible under *Kumho* (but not *Daubert*). She ultimately permitted the expert to identify similarities between two handwriting samples but not to draw conclusions on authorship (see *U.S. v. Hines*, 1999). The details of our summary were modeled after the original case, including the race and gender of the perpetrator and bank teller, the name and location of the bank, the amount of money stolen, and the bank teller's description of the perpetrator and subsequent eyewitness identification of Hines.

In our summary, a young African American man gave a handwritten note to a bank teller which read, "I have a gun. Keep quiet or I will shoot you. Give me all your cash!" The robber then opened his coat to reveal a concealed handgun and fled with over $10,000 in cash. Police arrived and interviewed the bank teller, who described the robber as a tall Black male wearing a heavy black coat and jeans, a general description that was consistent with surveillance footage.

Approximately a half hour after the robbery, police stopped a speeding vehicle in the vicinity of the bank, and found that the driver, Johanna Hines, generally matched the description given by the bank teller. Police reported that Hines appeared nervous while being interviewed, but a search of his vehicle revealed neither a gun nor the stolen cash. When shown a photo lineup of six individuals who fit the description that she had given, the bank teller identified Hines as the culprit but admitted that she was not 100% confident in her identification. At that point, Hines was picked up by police and brought to the station for questioning. Prior to the interview, Hines provided a handwritten waiver of his *Miranda* rights, which read, "I understand my rights to remain silent and to call a lawyer and I agree to talk at this time." Mr. Hines was then questioned for three hours.

Confession manipulation. Participants in the *confession-present* condition were told that after three hours of questioning Hines gave and signed a confession, which was shown to participants as a typed statement in which Hines stated that he robbed the bank because he was out of work and in debt. The confession statement was presented in typed form and signed illegibly to ensure that participants in the confession-present and confession-absent conditions were given the same amount of handwritten text to compare. He also described how he acquired and hid the gun, disguised his appearance, and hid the stolen cash somewhere so he could later retrieve it. After meeting with a lawyer, however, Hines

recanted his confession, claiming he was coerced by police, and pleaded not guilty.

Participants in the *confession-absent* condition were told that Hines maintained his innocence throughout the three-hour interview. When asked for his whereabouts during the robbery, Hines told police that he was eating breakfast alone at a local restaurant. He also provided the name and address of the restaurant and a description of what he had eaten for breakfast; these minutiae were included to ensure that Hines' confession and denial statements were roughly equivalent in terms of their level of detail. Despite his denials, Hines was charged with armed robbery, and pleaded not guilty.

Handwriting similarity manipulation. All participants were shown one of the four handwriting sample pairs that were previously pilot tested, all of which consisted of a robbery note and a *Miranda* waiver that were authored by different individuals. To investigate whether the impact of the confession was moderated by the perceived similarity of the handwriting samples, participants in the *high-similarity* condition received one of the two high-similarity pairs; those in the *low-similarity* condition received one of the two low-similarity pairs.

Dependent measures. Participants made a total of three judgments. First, they rated the similarity of the handwriting contained in the robbery note written by the perpetrator and the *Miranda* waiver written by Hines, the defendant. These ratings were made on a scale from 1 (*not at all similar*) to 10 (*very similar*). Second, they indicated whether they believed that the two notes were authored by the same individual ("actual match") and rated their confidence in that judgment on a scale from 1 (*not at all confident*) to 10 (*very confident*). Finally, they rendered a judgment of guilty or not guilty and rated their confidence in that judgment using the same 10-point scale.

Match and guilt judgments were measured in dichotomous response formats and their associated confidence ratings were made on a 10-point scale. This allowed for the construction of more sensitive continuous variables (ranging from −10 to + 10) by computing the product of the dichotomous judgment (coded as −1 or + 1) and confidence rating (1–10).

Comprehension test. After making these judgments, participants answered five multiple-choice questions designed to ensure that they read, understood, and remembered the contents of the case summary. This test included one critical item which asked whether the defendant had previously confessed to the robbery. Two participants were excluded after answering this item incorrectly.

Results

Similarity ratings. Of primary interest was the hypothesis that knowledge of a prior confession would lead participants to rate the defendant's handwriting as more similar to that of the perpetrator. Across conditions, participants generated a mean similarity rating of 4.17 ($SD = 2.30$). An independent-samples t test indicated that the differences in the mean ratings of the Confession-Present ($M = 4.47$, $SD = 2.21$) and Confession-Absent ($M = 3.86$, $SD = 2.36$) conditions were in the predicted direction. Although the effect was moderate in size, the difference did not achieve statistical significance, $t(167) = 1.73$, $p = .085$, $d = 0.27$, 95% CI [−0.07, 0.61].

To test for moderation by Handwriting Similarity, a 2 (Confession: Present vs. Absent) × 2 (Similarity: Low vs. High) factorial ANOVA was performed. Reiterating the results of the preceding test, the main effect of Confession approached but did not achieve significance, $F(1, 165) = 3.56$, $p = .061$. However, a main effect of Similarity confirmed the results of our pilot testing, $F(1, 165) = 17.65$, $p < .0001$, $d = 0.64$, 95% CI [0.31, 0.97], with high-similarity pairs ($M = 4.87$, $SD = 2.23$) being rated as more similar than low-similarity pairs ($M = 3.47$, $SD = 2.16$). The Confession × Similarity interaction was not significant, $F(1, 165) = 0.16$, $p = .691$, $\eta_p^2 = .001$.

Match judgments. We hypothesized that knowledge of a prior confession would increase the likelihood of participants incorrectly judging the two samples as an actual match—a binary judgment with serious consequences. Overall, only 18.93% of participants judged the two samples as a match. As predicted, however, participants in the Confession-Present condition judged the samples as a match significantly more often than those in the Confession-Absent condition (26.74% vs. 10.84%, respectively), $\chi^2(1) = 6.96$, $p = .008$, $OR = 3.00$, 95% CI [1.30, 6.96]. We next tested whether this effect was moderated by our handwriting similarity manipulation. Although the effect of the confession on match judgments was significant under conditions of Low Similarity, $\chi^2(1) = 6.63$, $p = .010$, $\varphi = 0.28$, 95% CI [0.07, 0.49], but not High Similarity, $\chi^2(1) = 1.70$, $p = .192$, $\varphi = 0.14$, 95% CI [−0.07, 0.36], a Breslow-Day test for the homogeneity of odds ratios revealed no evidence of significant moderation by Handwriting Similarity, $\chi^2(1) = 1.55$, $p = .214$ (see Figure 2).

To obtain a more sensitive test of confession and similarity main effects as well their interaction, a match-confidence composite score was created by combining match judgments with their associated confidence ratings. Participants who judged the samples as a match were given a positive score, and those who judged them as a nonmatch were given a negative score, thus creating a scalar variable ranging from −10 (highly confident "nonmatch" judgment) to +10 (highly confident "match" judgment). A 2 × 2 ANOVA on these match-confidence scores indicated that match judgments were infrequently made (across conditions, the overall mean was −4.38; $SD = 5.63$). Supporting our primary hypothesis, a main effect of confession indicated that participants in the Confession-Present condition had significantly higher composite scores ($M = -3.36$, $SD = 6.25$) than those in the Confession-Absent condition ($M = -5.43$, $SD = 4.73$), $F(1, 165) = 6.03$, $p = .015$, $d = 0.38$, 95% CI [−0.45, 1.21]. Neither the main effect of Similarity, $F(1, 165) = 2.55$, $p = .112$, $d = 0.24$, 95% CI [−0.60, 1.08], nor the Confession × Similarity interaction, $F(1, 165) = 0.11$, $p = .738$, $\eta_p^2 = .001$, was significant.

Guilt judgments. Overall, only 15.38% of participants judged the defendant as guilty. Consistent with earlier research, participants in the Confession-Present condition more often judged the defendant as guilty than those in the Confession-Absent condition (24.42% vs. 6.02%, respectively), $\chi^2(1) = 10.98$, $p = .001$, $OR = 5.04$, 95% CI [1.80, 14.11]. As with match judgments, we found that while the Confession effect was significant in the Low Similarity condition, $\chi^2(1) = 10.56$, $p = .001$, $\varphi = 0.35$, 95% CI [0.14, 0.56], but not in the High Similarity condition, $\chi^2(1) = 2.78$, $p = .095$, $\varphi = 0.18$, 95% CI [−0.03, 0.40], a Breslow-Day test indicated that the Handwriting Similarity manipulation did not signif-

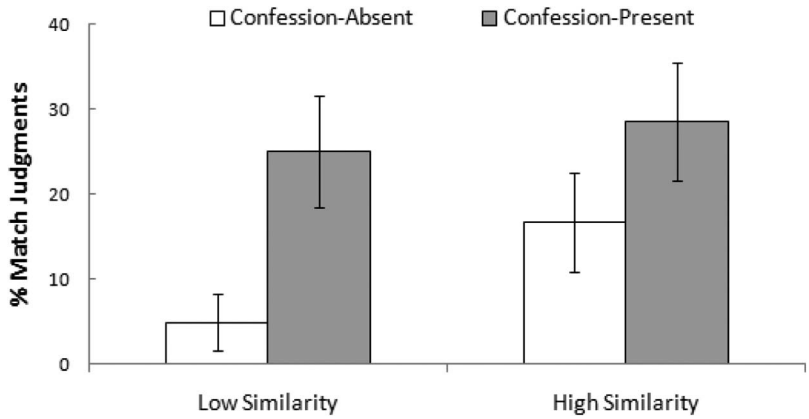


Figure 2. Effects of confession and handwriting similarity on match judgments (Study 1).

icantly moderate the effect of Confession on guilt judgments, $\chi^2(1) = 3.66$, $p = .056$ (see Figure 3).

As with match judgments, a guilt-confidence composite variable was created by combining guilt judgments with their associated confidence ratings. Consistent with the low proportion of guilty judgments, the mean guilt-confidence score across conditions was −4.81 ($SD = 5.46$). A 2 × 2 ANOVA on this measure revealed a significant main effect of confession, indicating that participants in the Confession-Present condition had significantly higher scores ($M = -3.48$, $SD = 6.27$) than those in the Confession-Absent condition ($M = -6.20$, $SD = 4.07$), $F(1, 165) = 11.43$, $p = .001$, $d = 0.52$, 95% CI [−0.28, 1.31]. Neither the main effect of Similarity, $F(1, 165) = 2.86$, $p = .093$, $d = 0.24$, 95% CI [−0.57, 1.06], nor the Confession × Similarity interaction, $F(1, 165) = 0.60$, $p = .440$, $\eta_p^2 = .004$, was significant.

Correlations between judgments of handwriting and guilt. Lastly, we assessed the links between the two sets of handwriting judgments obtained in response to our stimulus case and judgments of guilt. As anticipated, both similarity ratings, $r(167) = 0.52$, $p < .0001$, and match-confidence scores, $r(167) = 0.75$, $p < .0001$, were highly and significantly correlated with guilt-confidence scores. At least within the context of *Hines*, it is clear that perceptions of the handwriting stimuli were strongly linked to a belief in the defendant's guilt.

Study 2

Consistent with research in other domains of forensic science (e.g., Dror & Charlton, 2006; Elaad et al., 1994; Kassin, Bogart, & Kerner, 2012; Lange et al., 2011), the results of Study 1 suggest that knowledge of a confession can taint people's perceptions of handwriting evidence. When told that the defendant had confessed to a bank robbery—even though he went on to recant the confession, claiming it was coerced and false—participants were significantly more likely to conclude, erroneously, that the defendant had both authored the note and committed the robbery. In short, individuals who were told of a prior confession perceived the handwriting evidence as more incriminating than those who were not. Contrary to our hypothesis, these effects were not moderated by the prerated similarity of the handwriting samples.

The aim of Study 2 was to replicate our findings using a repeated-measures design in which participants judge the same pair of handwriting samples both before *and* after receiving case information. As others have noted (e.g., Dror et al., 2011),

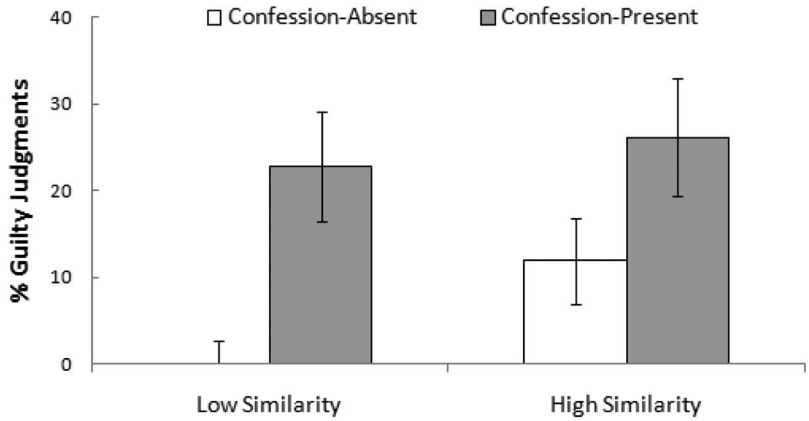


Figure 3. Effects of confession and handwriting similarity on guilt judgments (Study 1).

the use of a within-subjects, or intraobserver, design not only increases statistical power but allows us to more confidently draw conclusions regarding the biasing effect of the case information. Because we will compare participants' perceptions against their own prior judgments of the same stimuli, any observed effects cannot be attributed to individual differences in factors such as visual acuity, cognitive strategy, level of experience, and so forth. Moreover, demonstrating a lack of consistency in the same individuals over time would provide compelling evidence that these judgments are contingent on contextual (i.e., "top-down") factors such as the individual's mental set—not solely derived from the properties of the stimuli in question (i.e., "bottom-up" factors).

At Time 1, participants rated several pairs of handwriting samples without any contextual information. One week later, at Time 2, they read a version of the case summary used in Study 1 and were represented with a handwriting sample pair that they had previously evaluated. We hypothesized that the case information would cause their judgments of the same samples to change from Time 1 to Time 2, such that they would rate the samples as more similar in the presence of a confession and less similar in the absence of a confession. A secondary aim of Study 2 was to replicate our findings with a more diverse sample of individuals. Thus, we sought to obtain a nationally representative sample of eligible jurors by using the online marketplace service, Amazon Mechanical Turk (for a description, see Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Paolacci, Chandler, & Ipeirotis, 2010).

Method

Participants and design. A total of 254 individuals participated at Time 1. Upon recruitment, nine participants indicated that they were not current U.S. citizens and were excluded, leaving $N = 245$ participants at Time 1. Of these, 128 (52.24%) returned at Time 2, at which point they were randomly assigned to either the *confession-present* ($n = 42$), *confession-absent* ($n = 42$), or *no-context control* ($n = 44$) condition. Thus, the current study employed a 2 (Time 1 vs. Time 2) × 3 (Confession: Present, Absent, or Control) mixed factorial design.

Our final sample consisted of 128 participants, all of whom reported that they were current U.S. citizens. The sample included at least one participant from 33 of the 50 U.S. states, 75% females, with a mean age of 35.95 ($SD = 12.60$). Compared with Study 1, the sample was more diverse with respect to ethnicity (76.56% White, 7.03% Hispanic, 6.25% Black, 4.69% Asian, and 5.47% other) and educational background (33.59% did not hold a college degree, 58.59% held a college degree, and 7.81% held a graduate degree).

Participants who did not return at Time 2 ($n = 117$) did not differ from our final sample in terms of gender, $\chi^2(1) = 0.07$, $p = .789$, $OR = 1.08$, 95% CI [0.61, 1.92], or ethnicity, $\chi^2(5) = 1.42$, $p = .922$, $\varphi = 0.08$, 95% CI [−0.05, 0.20]. There was an age difference, $t(243) = 2.20$, $p = .029$, $d = 0.28$, 95% CI [−1.26, 1.82], such that individuals who did not return at Time 2 ($M = 32.48$, $SD = 12.07$) were somewhat younger than our final sample. There was also a difference in terms of education level, $\chi^2(2) = 10.78$, $p = .005$, $\varphi = 0.21$, 95% CI [0.08, 0.34], such that individuals who held a college degree were more likely to return at Time 2 than those who either did not hold a college degree or held a graduate-level degree.

Procedure. At Time 1, participants were recruited via Amazon Mechanical Turk (mTurk), an open marketplace service for recruiting individuals to complete online tasks. Empirical studies of mTurk have positively appraised the service for providing inexpensive and efficient access to samples that are more diverse than those obtained using traditional methods, while providing data of equal or greater quality (Buhrmester et al., 2011; Paolacci et al., 2010).

After volunteering to complete the online study, participants at Time 1 were redirected to a survey Web site, where they gave informed consent and answered basic demographic questions. Next, participants were shown eight pairs of handwriting samples, presented sequentially and in a randomized order. These samples consisted of the eight highest-rated nonmatching pairs from pilot testing. As in the pilot study, these pairs were presented without contextual information. Participants were asked to rate each pair for similarity and shared authorship. After completing Time 1, participants were instructed to return to the mTurk Web site in one week to access the second part of the study.

Participants completed Time 2 between 5 and 9 days after completing the first phase of the experiment ($M = 6.54$ days, $SD = 0.64$), and were randomly assigned to one of three conditions. Participants in the *confession-present* condition read the same case summary that was used in the confession-present condition of Study 1; participants in the *confession-absent* condition read the same case summary that was used in the confession-absent condition. Participants in the *control* condition did not receive any case information at Time 2.

Participants were then shown one pair of handwriting samples that they had previously rated at Time 1. All participants received the same pair (Pair 15), which was chosen because its mean similarity rating at Time 1 ($M = 4.31$, $SD = 2.58$) was closest to the overall grand mean across all eight pairs ($M = 4.31$). Participants again rated this pair for similarity and shared authorship. In the two experimental conditions, they also made a judgment of guilt and took a comprehension test to ensure that they understood the case.

Dependent measures. At Time 1, all participants rated the similarity of the handwriting between the two samples on a 10-point scale. Then they indicated whether they believed the two samples were a match and rated their confidence in this judgment on a 10-point scale. At Time 2, all participants made these same judgments a second time. Those in the confession-present and confession-absent conditions were also asked to make a guilt judgment and rate their confidence in that judgment on a 10-point scale.

Comprehension test. As in Study 1, participants in the two experimental conditions answered five multiple-choice questions designed to ensure that they read, understood, and remembered the contents of the case summary. All participants correctly reported that the defendant did (or did not) confess. Because participants in the control condition did not receive a case summary, they did not complete this comprehension test.

Results

Similarity ratings. Across conditions and time, participants generated a mean similarity rating of 4.57 ($SD = 2.47$). Our main hypothesis was that participants would rate the same handwriting samples as more similar when paired with a confession than when presented without context. A dependent-samples t test supported this hypothesis, $t(41) = 2.66$, $p = .011$, $d = 0.58$, 95% CI [−0.13, 1.28], with participants in the Confession-Present condition rating the same pair as more similar at Time 2 ($M = 5.24$, $SD = 2.34$) than they had at Time 1 ($M = 4.10$, $SD = 2.34$). In contrast, there were no significant changes over time in the Confession-Absent condition, (Time 1: $M = 4.02$, $SD = 2.29$; Time 2: $M = 4.29$, $SD = 2.06$), $t(41) = 0.64$, $p = .528$, $d = 0.15$, 95% CI [−0.47, 0.84], or in the Control condition (Time 1: $M = 4.70$, $SD = 2.78$; Time 2: $M = 5.02$, $SD = 2.77$), $t(43) = 0.73$, $p = .472$, $d = 0.16$, 95% CI [−0.66, 0.98] (see Figure 4).

A 2 (Time 1 vs. Time 2) × 3 (Confession: Present, Absent, or Control) mixed ANOVA was then performed on similarity ratings. An unanticipated main effect of Time emerged, $F(1, 125) = 5.41$, $p = .022$, $d = 0.29$, 95% CI [−0.13, 0.72], such that similarity ratings were significantly higher at Time 2 ($M = 4.85$, $SD = 2.43$) than at Time 1 ($M = 4.28$, $SD = 2.48$). Neither the main effect of Confession, $F(2, 125) = 1.42$, $p = .248$, $\eta_p^2 = .022$, nor the Time × Confession interaction, $F(2, 125) = 1.32$, $p = .271$, $\eta_p^2 = .021$, reached significance.

Match judgments. Across conditions and time, participants judged the samples as a match in 25% of their judgments. We hypothesized that participants would more often judge the same handwriting samples as a match when paired with a confession than when paired with a denial or without context. A chi-square test supported this hypothesis, McNemar $\chi^2(1) = 5.40$, $p = .035$, $\varphi = 0.36$, 95% CI [0.06, 0.66], such that participants in the Confession-Present condition were more likely to judge the pair as a match at Time 2 (35.71%) than they had at Time 1 (14.29%). There were no significant changes in match judgments over time in the Confession-Absent condition, McNemar $\chi^2(1) = 0.82$, $p = .549$, $\varphi = 0.14$, 95% CI [−0.16, 0.44], or in the Control condition, McNemar $\chi^2(1) = 0.33$, $p = .774$, $\varphi = 0.09$, 95% CI [−0.21, 0.38] (see Figure 5).

As in Study 1, match-confidence composite scores were created, which could range from −10 to +10, enabling us to test the interaction of Condition and Time. Across conditions and time, participants produced a mean composite score of −3.80 ($SD = 6.49$). A 2 × 3 mixed ANOVA on match-confidence composite scores revealed no main effect of Condition, $F(2, 125) = 1.61$, $p = .205$, $\eta_p^2 = .025$, and a nonsignificant effect of Time, $F(1, 125) = 3.25$, $p = .074$, $d = 0.22$, 95% CI [−0.93, 1.31], such that composite scores were somewhat higher at Time 2 ($M = -3.17$, $SD = 6.65$) than Time 1 ($M = -4.43$, $SD = 6.29$). Although the Time × Condition interaction was not quite significant, $F(2, 125) = 2.48$, $p = .087$, $\eta_p^2 = .038$, simple effects tests of our primary hypothesis indicated that composite scores increased over time in the Confession-Present condition (Time 1: $M = -5.17$, $SD = 5.58$; Time 2: $M = -1.90$, $SD = 7.06$), $t(41) = 2.51$, $p = .016$, $d = 0.55$, 95% CI [−1.58, 2.24], but not in the Confession-Absent, $t(41) = -0.48$, $p = .633$, $d = 0.11$, 95% CI [−1.68, 1.61], or Control, $t(43) = 0.95$, $p = .345$, $d = 0.20$, 95% CI [−1.96, 2.34], conditions.

To address the possibility that participants may have become more confident in their ability to discern matches at Time 2 as a result of their experience at Time 1, a dependent-samples t test was performed on match confidence ratings (irrespective of their dichotomous match judgments). This analysis revealed that confidence in match judgments did not change between Time 1 ($M = 7.38$, $SD = 2.09$) and Time 2($M = 7.09$, $SD = 1.90$), $t(127) = 1.51$, $p = .132$, $d = 0.19$, 95% CI [−0.17, 0.52].

Guilt judgments. After reading our case summary at Time 2, participants in the Confession-Present and Confession-Absent conditions gave a dichotomous guilt judgment, along with a confidence rating from 1–10 which was used to create a guilt-confidence composite score. Overall, 23.81% of participants in these conditions judged the defendant guilty at Time 2. A chi-square test indicated that participants in the Confession-Present condition were far more likely to judge the defendant as guilty than those in the Confession-Absent condition (42.86% vs. 4.76%, respectively), $\chi^2(1) = 16.80$, $p < .0001$, $OR = 15.00$, 95% CI [3.20, 70.39]. An independent t test indicated that participants in the Confession-Present condition ($M = -1.02$, $SD = 7.33$) had higher composite scores than the Confession-Absent condition ($M = -6.29$, $SD = 3.38$), $t(82) = 4.23$, $p < .0001$, $d = 0.94$, 95% CI [0.27, 2.14].

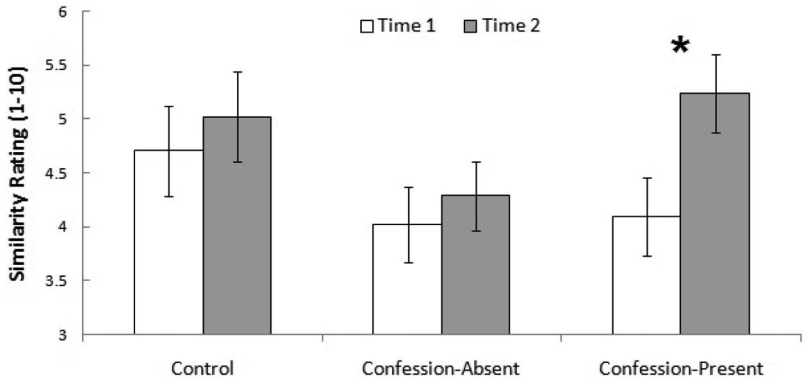


Figure 4. Effects of time and confession on similarity ratings (Study 2). * means differ at $p < .05$.

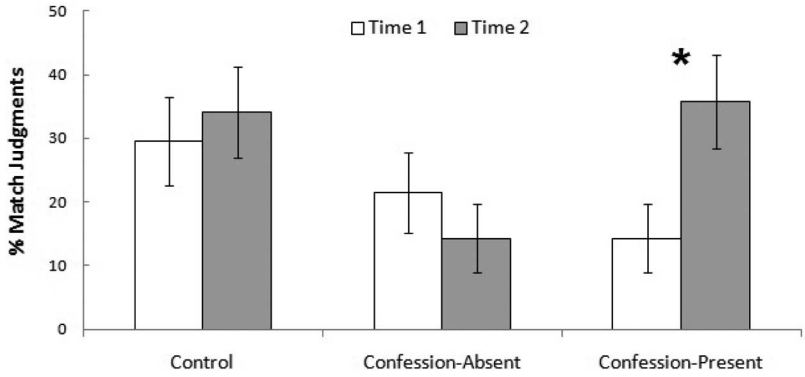


Figure 5. Effects of time and confession on match judgments (Study 2). * means differ at $p < .05$.

Correlations between judgments of handwriting and guilt. As in Study 1, we examined the associations between participants' judgments of handwriting and guilt in our stimulus case. Both similarity ratings, $r(82) = 0.63$, $p < .0001$, and match-confidence scores, $r(82) = 0.83$, $p < .0001$, were again strongly correlated with guilt-confidence scores, thus reiterating the interrelatedness of participants' judgments of handwriting and guilt in our stimulus case.

Summary. Study 2 replicated the main findings of Study 1 using a repeated-measures design in which participants rated the same pair of handwriting samples twice—the second time in the context of a criminal trial. When represented as evidence in a case in which the defendant had previously confessed (but then retracted his confession), participants rated the samples as more similar, and more often misjudged them to be a match, than they had earlier. In contrast, participants' judgments of handwriting stimuli did not change after being given the same case information without the confession. In short, participants with knowledge of a prior confession saw the handwriting evidence as more inculpatory, and were more likely to judge the defendant as guilty, compared with those who were told that the defendant had maintained his innocence.

Study 2 also produced an unanticipated effect of time on similarity ratings, such that participants judged the same handwriting samples as more similar when seeing them for a second time. We suggest two nonmutually exclusive explanations for this finding. First, though we predicted that confession-absent participants would infer innocence and thus provide lower similarity ratings at Time 2, it is possible that our confession-absent condition instead created a weak expectation of guilt which produced the observed nonsignificant increase in similarity ratings over time in that condition. Second, research on the phenomenon of repetition priming finds that individuals tend to process previously seen visual stimuli more efficiently, as attention is automatically drawn to previously attended-to features of the stimulus deemed important to the task at hand (for a review, see Kristjansson & Campana, 2010). Thus, if participants ascribed greater importance to similarities than dissimilarities when comparing handwriting samples at Time 1, they would be more apt to focus on these similarities at Time 2. In any event, the increase in perceived similarity over time reached significance only when a confession was present.

Discussion

In a series of studies, we tested the hypothesis that knowledge of a prior confession would taint people's perceptions and judgments of handwriting evidence. Pilot data indicated that people were willing to believe two handwriting samples could have been written by the same author even when they rated the samples as dissimilar, suggesting that their judgments were malleable and likely sensitive to context. Study 1 found support for this proposition; participants who read a case summary in which the defendant had previously confessed were more likely to conclude, erroneously, that the samples were authored by the same individual (i.e., a "match") and were more likely to judge the defendant as guilty, compared with a group that read the same case summary without a prior confession. Study 2 replicated these findings. Using a repeated-measures design, we found that participants rated the same handwriting samples as more similar, were more likely to judge them as a "match," and were more likely to judge the defendant as guilty when the samples were paired with a confession than when they were not. Illustrating the power of confession, it is notable that this manipulation proved potent even though the confession was relatively weak by virtue of retraction, a lack of detail indicating guilty knowledge, and the defendant's claim that it was coerced and false.

Consistent with basic psychological research on confirmation biases (see Nickerson, 1998), the current studies contribute to a burgeoning literature indicating that judgments of forensic science evidence can be shaped by the knowledge and expectations of the observer (e.g., Dror & Charlton, 2006; Dror et al., 2006; Miller, 1984) and that confessions in particular can guide the collection and evaluation of evidence so as to cohere with the confession (Elaad et al., 1994; Hasel & Kassin, 2009; Kassin et al., 2012). Our findings suggest the possibility that jurors who are presented with handwriting evidence at trial may be tainted in their evaluations of that stimulus information by their a priori belief in the defendant's guilt or innocence, which will have been shaped by their knowledge of other aspects of the case (e.g., a prior confession). It is noteworthy that the presence of a confession did not merely affect participants' dichotomous judgments that the defendant had written the note used in the robbery; it also impacted how much similarity they perceived between the two handwriting samples. Thus, it appears that they did not mindlessly acquiesce to the

confession's implication that the two samples would match, but instead their belief in the defendant's guilt affected their fundamental perception of the handwriting stimuli.

Our findings are consistent with Simon's (2004, 2011) *cognitive coherence* framework, which posits that legal decision making entails movement from complexity (i.e., conflicting items of evidence that must be reconciled and integrated) to coherence (i.e., a binary verdict supported by a bulk of the available evidence). Coherence is reached via a bidirectional process in which evidence leads the decision maker toward a conclusion, which concurrently shapes the evaluation of other evidence such that it will cohere with the emerging conclusion. This process was most evident in Study 2, where knowledge of a confession both cultivated an emerging belief in the defendant's guilt and led participants to adjust their own prior evaluations of handwriting samples. As a result, individual items of evidence become nonindependent. As Simon (2011) notes, an item that is considered probative (e.g., a confession) tends to bring other ambiguous evidence into coherence with it, thereby causing the ambiguous evidence as well to be seen as probative.

Accordingly, our findings should not be taken to suggest that it is irrational for assessments of handwriting evidence to be tainted by a confession; indeed it is quite rational for jurors to integrate all available information in forming holistic judgments of guilt. This phenomenon is problematic, however, insofar as it promotes the illusion among fact-finders that they have weighed the value of each item independently, when in fact they did not. Though we did not assess participants' awareness of the impact of the confession on their handwriting evaluations, others have found that individuals are often unaware of the extent to which their judgments are influenced by the larger evidentiary context. For example, Charman, Gregory, and Carlucci (2009) found that mock investigators' judgments of a facial composite were biased by their knowledge of inculpatory eyewitness evidence even when they claimed that the eyewitness did not influence them. Similarly, jurors may unwittingly devalue some items of evidence and overvalue others in pursuit of coherence, while also maintaining a naïve faith in their own objectivity (Simon, 2012).

Interdependence among discrete evidentiary judgments can result in what Kassin (2012) has called *corroboration inflation*. When knowledge of a prior confession causes ambiguous handwriting evidence to be viewed as incriminating, the confession has effectively created a second source of inculpatory evidence, giving the illusion that the strength of evidence against the defendant is stronger (and more coherent) than it truly is. As Simon (2012) explains, "evidence that might otherwise have given rise to a reasonable doubt can be reduced in the fact finder's mind to a mere negligible doubt" (p. 175). Insofar as confessions carry a uniquely strong implication of guilt (e.g., Kassin & Neumann, 1997), they may be especially likely to corrupt other evidence and create corroboration inflation. Yet this effect is likely not limited to confessions, as other types of incriminating information, notably eyewitness identifications, can also corrupt evidentiary judgments—a notion that has received some empirical support as well (Charman et al., 2009). Even nonprobative information, such as exposure to gruesome crime scene photos, has been shown to affect judgments of forensic evidence in some cases (Dror, Peron, Hind, & Charlton, 2005). In short, mounting evidence suggests that legal decision makers do not evaluate multiple items of information independently of each other; instead, their integration entails a complex interdependence between each of the individual items, such that the body of evidence as a whole can become greater than the sum of its parts (see also Charman, 2013).

In theory, any type of information can impact one's judgments of any other type of information during the evidence integration process. Thus, although participants evaluated handwriting evidence differently when they were aware of a prior confession, their appraisal of the handwriting may have likewise affected how much credence they gave to the confession. In both studies, a minority of participants judged the defendant as guilty despite being told that he had confessed (24.42% in Study 1; 42.86% in Study 2). These numbers are unusually low in comparison to typical conviction rates when mock jurors are told of a confession (e.g., Kassin & Neumann, 1997; Kassin & Sukel, 1997). In part, this may be due to the fact that we presented a rather weak confession: it contained no narrative details that betrayed guilty knowledge, it was obtained under somewhat coercive circumstances, and was recanted shortly thereafter. In addition, it is possible that the handwriting evidence weakened the impact of the confession as part of the process of evidence integration. Overall, participants in the confession conditions rated our handwriting samples as only moderately similar (Ms = 4.47 and 5.24 on a 10-point scale, in Studies 1 and 2, respectively). To the extent that dissimilarities between the two samples suggested the defendant's innocence, participants who were also faced with a confession may have reconciled this seeming contradiction by attributing less probative value to the confession.

The question of how to prevent interevidentiary influences in juror decision making is not easily answered. One possible solution is for judges to administer curative instructions to juries warning against the possibility that one item of evidence can taint their assessments of others. Meta-analytic evidence suggests that the effectiveness of such admonishment is limited (Steblay, Hosch, Culhane, & McWethy, 2006), particularly with regards to evidence that is seen as reliable and probative of guilt (e.g., Kassin & Sommers, 1997; Sommers & Kassin, 2001). As we did not specifically instruct participants to disregard their knowledge of other case facts when evaluating the handwriting evidence, this question remains an important subject of future research.

Given that ambiguity is conducive to the operation of confirmation biases (e.g., Kunda, 1990), Study 1 also tested the hypothesis that the effect of the confession on handwriting judgments would be moderated by the prerated similarity of the samples being compared. Others have previously found such effects—for example, biasing case information impacted how polygraph examiners scored inconclusive, but not conclusive, charts (Elaad et al., 1994) and had a greater effect on latent fingerprint expert judgments of "relatively difficult," as opposed to "not difficult," stimuli (Dror & Charlton, 2006). However, Study 1 found no significant evidence for moderation by similarity.

One explanation for this null effect is that neither the "low similarity" pair nor the "high similarity" stimulus pair preempted a subjective analysis. Kunda (1990) noted that *reality constraints* inhibit the effects of confirmation bias: When two stimuli are clearly different or identical, even very strong expectations to the contrary do not enable the observer to distort the sensory input to a degree that validates their incongruent prior belief. It is possible, then, that our low similarity samples were not sufficiently dissim-

ilar, nor were our high similarity samples sufficiently similar, to preempt the operation of confirmation bias. A second, related possibility is that participants possessed an intrinsic sensitivity to intraauthor variability in handwriting style, such that no pair of samples would be readily seen as an "obvious" match or nonmatch. Indeed, some have suggested that natural variability in a person's handwriting over time renders the task of handwriting identification virtually impossible—even for trained experts (Risinger & Saks, 1996). This latter explanation is also consistent with pilot data showing that people believed that dissimilar samples could have nonetheless come from the same author.

Although the current studies tested laypeople rather than professional handwriting examiners, our findings also underscore concerns raised by the National Academy of Sciences (2009) regarding the potential for cognitive bias to compromise the validity of forensic science judgments. It appears that professional forensic examiners often receive extraneous information that could influence their analysis—and that this information can bias their interpretations of stimulus samples (Kassin et al., 2013). Indeed, Risinger, Saks, Thompson, and Rosenthal (2002) characterize the practice of submitting case information along with evidence to be analyzed as "virtually universal" (p. 32). A growing body of research suggests that this knowledge can sway the judgments of even experienced examiners, as in one study where 17% of latent fingerprint experts' judgments changed in the face of contextual information (Dror & Charlton, 2006). It is important to emphasize that this phenomenon is not necessarily attributable to examiner negligence, incompetence, or misconduct (Dror, Kassin, & Kukucka, 2013). Rather, due to the pervasive "human nature" of confirmation biases (Klayman & Ha, 1987; Nickerson, 1998), it follows that even trained and well-intentioned experts can unknowingly be influenced by biasing information. Additional research is sorely needed across various domains of forensic science to determine the effects of expertise and training on people's susceptibility to confirmation bias. As there exists only one empirical study on the subject of bias among forensic handwriting examiners (Miller, 1984), we hope that the paradigm used in the current studies can be adapted for use with this population.

Future studies of bias among forensic examiners would also have implications for the admissibility of their testimony as experts. Handwriting examiners are frequently proffered as experts at trial (Risinger, 2007). Although their testimony may not meet the *Daubert* standard for reliable scientific knowledge, it has been admitted as nonscientific specialized knowledge under *Kumho* (e.g., *U.S. v. Hines*, 1999). Given the persuasive impact of forensic science testimony on jurors (McQuiston-Surrett & Saks, 2009) and the discovery of invalid forensic science testimony in a large proportion of DNA exoneration cases (Garrett & Neufeld, 2009), it is imperative that forensic experts' conclusions be based solely from the physical evidence in question. Studies of expert performance can be informative in this regard. If data indicate that expert judgments are highly accurate and immune to confirmation bias, this will strengthen their case for admission under *Daubert*; conversely, if their judgments are shown to be tainted by contextual information, such testimony could mislead rather than assist the trier of fact.

Although it remains an empirical question concerning the extent to which handwriting experts are vulnerable to contextual bias, at least one forensic laboratory has taken preemptive measures to protect against the pernicious possibilities. Found and Ganas (in press) describe an ongoing program of procedural changes in an Australian forensic handwriting laboratory designed to reduce examiners' exposure to potentially biasing information. Echoing advocates of context-blind testing (e.g., Kassin et al., 2013), these practitioners argue that isolating examiners from irrelevant information should be the default practice to best minimize the risks of bias and error. To that end, they reviewed the case information sheet that accompanies their receipt of evidence and eliminated all fields containing information that they agreed was unnecessary for the analysis—which notably included "any description of any admissions made in relation to the case." Over 3 years later, Found and Ganas (in press) reported that the procedural changes they enacted were not "overly time-consuming or expensive," produced a number of advantageous outcomes, and had no negative ones. This account stands in stark contrast to critics of blind testing who have argued that reforms to shield examiners from extraneous information could be prohibitively expensive (Charlton, 2013) and detract from examiners' accuracy (Butt, 2013; Elaad, 2013).

In addition to context-blind testing, other reforms have been proposed to mitigate the possible effects of confirmation bias on forensic examiners. For example, Saks, Risinger, Rosenthal, and Thompson (2003) proposed that an intermediary should serve as the sole contact point between investigators and examiners, allowing them to filter out extraneous information (i.e., "masking") and restrict other suggestive communications between the two parties (see also Risinger, 2009). Others (e.g., NAS, 2009; Risinger et al., 2002) suggested that asking examiners to test evidence from only one suspect may be suggestive, as examiners may infer guilt from the base-rate assumption that investigators did not choose this suspect at random. The proposed solution is the use of an "evidence lineup," where forensic examiners are given an array of suspect samples and asked to determine which, if any, matches the crime-relevant sample. Proponents of evidence lineups note that they would protect against contextual bias and allow for the estimation of error rates (Wells, Wilford, & Smalarz, 2013), their construction and administration could be informed by the eyewitness psychology literature (Kassin et al., 2013), and they would not be financially burdensome, especially if the lineups are constructed by the same intermediary who handles the masking procedures (Reese, 2012). Despite frequent discussion, there exists only one empirical study of evidence lineups. Miller (1987) found that students trained in human hair identification made fewer incorrect judgments when comparing a crime scene hair against a lineup of five nonmatching hairs rather than a single nonmatching hair. This study did not, however, specifically address whether evidence lineups protect against bias. Given the promise of this measure of reform, future research should explore the effects of evidence lineup use on examiners' accuracy and bias.

In sum, in light of the alarming frequency with which forensic science errors have been implicated as contributing factors in known DNA exoneration cases (Hampikian et al., 2011; www.innocenceproject.org), the National Academy of Sciences (2009) lamented that "research has been sparse on the important topic of cognitive bias in forensic science—both regarding their effects and methods for minimizing them" (p. 124). The current studies advance our understanding of the process whereby forensic confirmation bias can contribute to wrongful convictions. Although these findings constitute a valuable first step, additional studies are

needed to identify the conditions that promote bias as well as protective reforms aimed at minimizing its detrimental impact.

References

Ask, K., Rebelius, A., & Granhag, P. A. (2008). The "elasticity" of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology, 22,* 1245–1259. doi:10.1002/acp.1432

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91,* 612–625. doi:10.1037/0022-3514.91.4.612

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science, 21,* 147–152. doi:10.1177/0956797609356283

Boring, E. G. (1930). A new ambiguous figure. *The American Journal of Psychology, 42,* 444–445. doi:10.2307/1415447

Bressan, P., & Dal Martello, M. F. (2002). "Talis pater, talis filius:" Perceived resemblance and the belief in genetic relatedness. *Psychological Science, 13,* 213–218. doi:10.1111/1467-9280.00440

Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science, 144,* 424–425. doi:10.1126/science.144.3617.424

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Butt, L. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions: Commentary by a forensic examiner. *Journal of Applied Research in Memory & Cognition, 2,* 59–60.

Charlton, D. (2013). Standards to avoid bias in fingerprint examination: Are such standards doomed to be based on fiscal expediency? *Journal of Applied Research in Memory & Cognition, 2,* 71–72.

Charman, S. D. (2013). The forensic confirmation bias: A problem of evidence integration, not just evidence evaluation. *Journal of Applied Research in Memory & Cognition, 2,* 56–58.

Charman, S. D., Gregory, A. H., & Carlucci, M. (2009). Exploring the diagnostic utility of facial composites: Beliefs of guilt can bias perceived similarity between composite and suspect. *Journal of Experimental Psychology: Applied, 15,* 76–90. doi:10.1037/a0014682

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon Mechanical Turk as a tool for experimental behavioral research. *PLoS one, 8,* e57410. doi:10.1371/journal.pone.0057410

Daubert v. Merrell. (1993). Dow Pharmaceuticals, 509, U.S. 579.

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-author consistency and the effect of a "target" comparison. *Forensic Science International, 208,* 10–17. doi:10.1016/j.forsciint.2010.10.013

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification, 56,* 600–616.

Dror, I. E., Charlton, D., & Peron, A. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156,* 74–78. doi:10.1016/j.forsciint.2005.10.017

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice, 51,* 204–208. doi:10.1016/j.scijus.2011.08.004

Dror, I. E., Kassin, S. M., & Kukucka, J. (2013). New application of psychology to law: Improving forensic evidence and expert witness contributions. *Journal of Applied Research in Memory & Cognition, 2,* 78–81.

Dror, I. E., Peron, A. E., Hind, S.-L., & Charlton, D. (2005). When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology, 19,* 799–809. doi:10.1002/acp.1130

Dunning, D., & Balcetis, E. (2013). Wishful seeing: How preferences shape visual perception. *Current Directions in Psychological Science, 22,* 33–37. doi:10.1177/0963721412463693

Elaad, E. (2013). Psychological contamination in forensic decisions. *Journal of Applied Research in Memory & Cognition, 2,* 76–77.

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7,* 279–292. doi:10.1002/bdm.3960070405

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review, 2,* 291–397.

Found, B., & Ganas, J. (in press). The management of domain irrelevant context information in forensic handwriting examination casework. *Science & Justice*.

Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97.

Halverson, A. M., Hallahan, M., Hart, A. J., & Rosenthal, R. (1997). Reducing the biasing effects of judges' nonverbal behavior with simplified jury instruction. *Journal of Applied Psychology, 82,* 590–598. doi:10.1037/0021-9010.82.4.590

Hampikian, G., West, E., & Akselrod, O. (2011). The genetics of innocence: Analysis of 194 U.S. DNA exonerations. *Annual Review of Genomics and Human Genetics, 12,* 97–120. doi:10.1146/annurev-genom-082509-141715

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 20,* 122–126. doi:10.1111/j.1467-9280.2008.02262.x

Henkel, L. A., Coffman, K. A. J., & Dailey, E. M. (2008). A survey of people's attitudes and beliefs about false confessions. *Behavioral Sciences & the Law, 26,* 555–584. doi:10.1002/bsl.826

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology, 13,* 357–371. doi:10.1348/135532507X238682

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

Kam, M., Fielding, G., & Conn, R. (1997). Writer identification by professional document examiners. *Journal of Forensic Sciences, 42,* 778–786. doi:10.1520/JFS14207J

Kam, M., & Lin, E. (2003). Writer identification using hand-printed and non-hand-printed questioned documents. *Journal of Forensic Sciences, 48,* 1391–1395.

Kassin, S. M. (2012). Why confessions trump innocence. *American Psychologist, 67,* 431–445. doi:10.1037/a0028212

Kassin, S. M., Bogart, D., & Kerner, J. (2012). Confessions that corrupt: Evidence from the DNA exoneration case files. *Psychological Science, 23,* 41–45. doi:10.1177/0956797611422918

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory & Cognition, 2,* 42–52.

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27,* 187–203. doi:10.1023/A:1022599230598

Kassin, S. M., & Neumann, K. (1997). On the power of confession evidence: An experimental test of the "fundamental difference" hypothesis. *Law and Human Behavior, 21,* 469–484. doi:10.1023/A:1024871622490

Kassin, S. M., & Sommers, S. R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality and Social Psychology Bulletin, 23,* 1046–1054. doi:10.1177/01461672972310005

Kassin, S. M., & Sukel, H. (1997). Coerced confessions and the jury: An experimental test of the "harmless error" rule. *Law and Human Behavior, 21,* 27–46. doi:10.1023/A:1024814009769

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Bulletin, 94,* 211–228. doi:10.1037/0033-295X.94.2.211

Kristjansson, A., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, Psychophysics, 72,* 5–18. doi:10.3758/APP.72.1.5

Kumho Tire Co. v. Carmichael, 526, U.S. 137 (1999).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. doi:10.1037/0033-2909.108.3.480

Lange, N. D., Thomas, R. P., Dana, J., & Dawes, R. M. (2011). Contextual biases in the interpretation of auditory evidence. *Law and Human Behavior, 35,* 178–187. doi:10.1007/s10979-010-9226-4

Leeper, R. (1935). A study of a neglected portion of the field of learning: The development of sensory organization. *The Pedagogical Seminary and Journal of Genetic Psychology, 46,* 41–75. doi:10.1080/08856559.1935.10533144

Leo, R. A., & Liu, B. (2009). What do potential jurors know about police interrogation techniques and false confessions? *Behavioral Sciences & the Law, 27,* 381–399. doi:10.1002/bsl.872

Lynch, M. (2003). God's signature: DNA profiling, the new gold standard in forensic evidence. *Endeavor, 27,* 93–97. doi:10.1016/S0160-9327(03)00068-1

McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and Human Behavior, 33,* 436–453. doi:10.1007/s10979-008-9169-1

Miller, L. S. (1984). Bias among forensic document examiners: A need for procedural changes. *Journal of Police Science and Administration, 12,* 407–411.

Miller, L. S. (1987). Procedural bias in forensic science examinations of human hair. *Law and Human Behavior, 11,* 157–163. doi:10.1007/BF01040448

Mnookin, J. L., Cole, S. A., Dror, I. E., Fisher, B. A. J., Houck, M. M., Inman, K., . . . Stoney, D. A. (2011). The need for a research culture in the forensic sciences. *UCLA Law Review, 58,* 725–779.

Narchet, F. M., Meissner, C. A., & Russano, M. B. (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law and Human Behavior, 35,* 452–465. doi:10.1007/s10979-010-9257-x

National Academy of Sciences. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. doi:10.1037/1089-2680.2.2.175

O'Brien, B. (2009). Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy, and Law, 15,* 315–334. doi:10.1037/a0017881

Paolacci, J., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419. doi:10.2139/ssrn.1626226

Reese, E. J. (2012). Techniques for mitigating cognitive biases in fingerprint identification. *UCLA Law Review, 59,* 1252–1290.

Risinger, D. M. (2007). Cases involving the reliability of handwriting identification expertise since the decision in *Daubert*. *Tulsa Law Review, 43,* 477–596.

Risinger, D. M. (2009). The NAS report on forensic science: A glass nine-tenths full (this is about the other tenth). Published online July 21, 2009 at the Social Science Research Network: http://ssrn.com/abstract=1437276

Risinger, D. M., & Saks, M. J. (1996). Science and nonscience in the courts: Daubert meets handwriting identification expertise. *Iowa Law Review, 82,* 21–74.

Risinger, D. M., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review, 90,* 1–56. doi:10.2307/3481305

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science, 309,* 892–895. doi:10.1126/science.1111565

Saks, M. J., Risinger, D. M., Rosenthal, R., & Thompson, W. C. (2003). Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. *Science & Justice, 43,* 77–90. doi:10.1016/S1355-0306(03)71747-X

Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *The University of Chicago Law Review, 71,* 511–586.

Simon, D. (2011). The limited diagnosticity of criminal trials. *Vanderbilt Law Review, 64,* 143–223.

Simon, D. (2012). *In doubt: The psychology of the criminal justice process*. Cambridge, MA: Harvard University Press. doi:10.4159/harvard.9780674065116

Sommers, S. R., & Kassin, S. M. (2001). On the many impacts of inadmissible testimony: Selective compliance, need for cognition, and the overcorrection bias. *Personality and Social Psychology Bulletin, 27,* 1368–1377. doi:10.1177/01461672012710012

Steblay, N., Hosch, H. M., Culhane, S. E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior, 30,* 469–492. doi:10.1007/s10979-006-9039-7

U.S. v. Buck, 1987 WL 19300 (S. D. N. Y.). (1987).

U.S. v. Hines, 55, F.Supp.2d 62 (1999).

U.S. v. Paul, 175 F. 3d 906, (1999).

U.S. v. Starzecpyzel, 880 F. Supp. 1027 (1995).

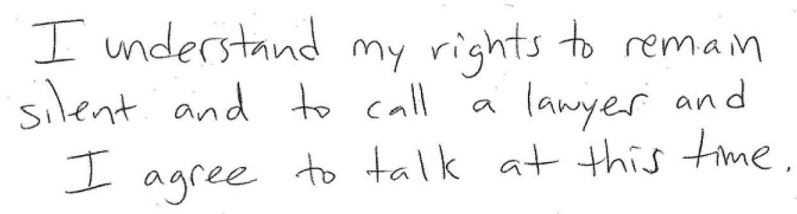
Wells, G. L., Wilford, M. M., & Smalarz, L. (2013). Forensic science testing: The forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports. *Journal of Applied Research in Memory & Cognition, 2,* 53–55.

(*Appendix follows*)

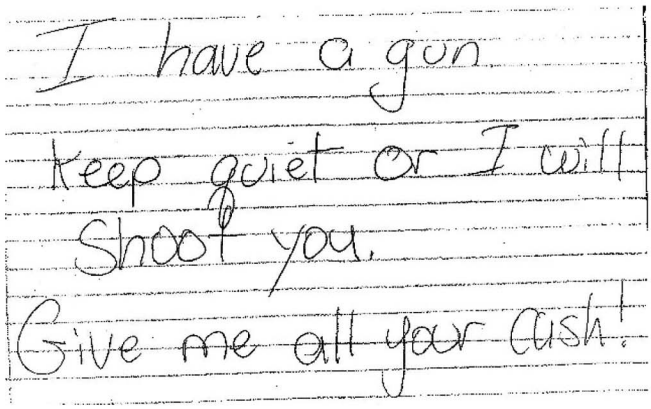
Appendix

Examples of Pilot-Tested Handwriting Stimuli Selected for Use in Study 1

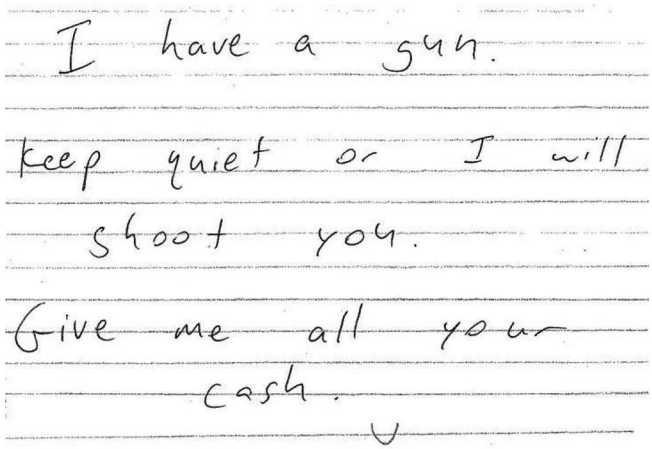
Target sample (constant across all conditions)



Example of a low-similarity comparison sample (from Pair #6)



Example of a high-similarity comparison sample (from Pair #14)



Received March 24, 2013
Revision received September 5, 2013
Accepted September 7, 2013 ■