Law and Human Behavior
2016, Vol. 40, No. 1, 50–64

0147-7307/16/$12.00 http://dx.doi.org/10.1037/lhb0000154

Putting Bias Into Context: The Role of Familiarity in Identification

Rachel A. Searston and Jason M. Tangen
The University of Queensland

Kevin W. Eva
The University of British Columbia

Previous demonstrations of context effects in the forensic comparison sciences have shown that the number of "match" responses a person makes can be swayed by case information. Less clear is whether these effects are a result of changes in *accuracy* (e.g., discrimination ability), a shift in *response bias* (e.g., tendency to say "match" or "no match") or a mix of the 2. We present a series of experiments where we use a signal detection framework to examine the effects of case information (separately) on forensic comparison accuracy and response bias. We also explore the role of familiarity as 1 potential mechanism for case information to sway accuracy. In Experiment 1, case information about crimes perceived to be more severe swayed people to say "match" more, but had little bearing on their ability to discriminate matching and nonmatching fingerprint pairs. In Experiment 2, case information did affect accuracy when it was familiar (i.e., if a previous similar case was associated with a "match" then people were more likely to also rate the current case as a "match," even though it was not). Even when we blinded people to all extrinsic case information in Experiment 3, accuracy was significantly affected by the familiarity of the fingerprints. These results demonstrate that contextual factors can have different (and independent) influences on accuracy and response bias and that even subtle information can affect accuracy if it is sufficiently similar to the case or trace at hand.

Keywords: cognitive bias, context effects, identification, instance-based learning, familiarity

Forensic science practices have been under review, and heavily criticized by several prominent scientific bodies. In 2009, the U.S. National Academy of Sciences issued a comprehensive report about the state of practices in forensic science, suggesting that sources of bias and human error are likely to contribute to wrongful arrests of innocent people (National Research Council, 2009; see also Garrett & Neufeld, 2009). Similar concerns regarding human error and lack of a research culture within the forensic science community have since been raised in reports issued by the Scottish Public Judicial Inquiry into fingerprinting (Campbell, 2011) and the National Institute of Standards and Technology Expert Working Group on Human Factors in Latent Print Analysis (2012). The National Academy of Sciences report specifically included a recommendation for the establishment of ". . . research programs on human observer bias and sources of human error in forensic examinations," along with the recommendation that such programs would benefit from drawing on established findings in diagnostic medicine and cognitive psychology (National Research Council, 2009, S-18).

Here, we critically review the existing research on context effects in the forensic sciences. We then draw on research in cognitive psychology to offer an instance-based account of these contextual influences. Finally, we present three experiments to better understand the extent to which specific contextual factors affect human performance (response bias and accuracy) in forensic comparison tasks.

Context Effects in Forensic Science

The question of how much case information an examiner ought to have at her disposal is a hot topic in forensic science (Champod, 2014). Some commentators insist that knowledge of the case could potentially influence an examiner's judgment (e.g., overweighting the degree of similarity between a pair of fingerprints in light of a confession) and that they should not have access to particular aspects of a case (e.g., see Kassin, Dror, & Kukucka, 2013, for review and recommendations). Others suggest that examiners require access to certain bits of information about the case to make an informed decision (e.g., Butt, 2013; Champod, 2014) and that blinding procedures, if strictly adhered to, create the risk of dedicating finite resources that could be better spent elsewhere.

Across academic forums, ideas and strategies about how best to minimize and control for context effects are also at the forefront of discussions. A recent review of the forensic confirmation bias literature by Kassin et al. (2013) sparked commentary from forensic practitioners and academics on the recommendations made by the authors to remove extraneous case information in forensic laboratories. Kassin et al. (2013) made several recommendations for reforming practices in forensic science. For example, examiners should complete and document their analyses of the trace evidence in isolation prior to making a comparison to known targets; blind and double-blind procedures ought to be implemented and strictly adhered to throughout the identification process (i.e., restricting communication with the investigator or removing details about the case, including conclusions drawn by

This article was published Online First August 24, 2015.

Rachel A. Searston and Jason M. Tangen, School of Psychology, The University of Queensland; Kevin W. Eva, Centre for Health Education Scholarship, The University of British Columbia.

Correspondence concerning this article should be addressed to Rachel A. Searston, School of Psychology, University of Queensland, St. Lucia, Queensland, Australia 4072. E-mail: rachel.searston@gmail.com

previous examiners); cross laboratory verification should be used wherever possible; candidate lists should be presented randomly to examiners; and certification and training in forensic science ought to include requirements for a basic understanding of the experimental method, perception, and decision-making.

In response, others highlighted concerns regarding the fiscal (Charlton, 2013) and practical (Butt, 2013) costs of full-scale implementation of the proposed recommendations in forensic laboratories that are already stretched for time, staff, and resources. Other commentators, including Dror, Kassin, and Kukucka (2013) in their reply, argued that many of the recommendations are relatively low cost (Cole, 2013) and may even improve the efficiency of examiners' workflow (e.g., by eliminating the time taken on irrelevant tasks such as reading extraneous case information). Several commentators emphasized the need for academic and professional stakeholders to work collaboratively in developing new and applied research programs and to seek mutually agreeable solutions to managing sources of bias (Charlton, 2013; Haber & Haber, 2013; Heyer & Semmler, 2013).

Overall, there appears to be a consensus that the potential for contextual information to influence forensic analyses is a concern, and that instituting strategies to safeguard against these influences would improve the quality and reliability of forensic evidence (Cole, 2013; Dror et al., 2013; Wells, Wilford, & Smarlarz, 2013). Less is known, however, about the nature of the problem: Does the context of a case influence accuracy? Does it influence a person's tendency to make a false alarm over a miss error? And what are the cognitive mechanisms driving these effects? We aim to address some of this uncertainty in the present series of experiments.

The Research Base

The research base on contextual influences and human error in the forensic sciences has grown recently, but is still in its early stages. For more than a century, fingerprint evidence has been considered irrefutable in forensic science (Cole, 2004). Until very recently, there has been no good measurement of the accuracy of human fingerprint examiners at all. The data are now clear, however, that human fingerprint examiners are incredibly accurate compared to novices at comparing fingerprints (e.g., Tangen, Thompson, & McCarthy, 2011; Thompson, Tangen, & McCarthy, 2014). We have also learned that examiners are fallible, and tend to err on the side of caution by preferring to make errors of the sort that would fail to identify a criminal (misses) rather than provide evidence to the court that would incorrectly convict an innocent person (false alarms; Tangen et al., 2011; Thompson et al., 2014; Ulery, Hicklin, Buscaglia, & Roberts, 2011).

A few experiments, mainly by Itiel Dror and his colleagues, have examined the influence of contextual information on human interpretation of forensic evidence (e.g., Dror, Charlton, & Péron, 2006; Dror, Péron, Hind, & Charlton, 2005; see Kassin et al., 2013, for a review). These studies provide evidence that expert examiners make decisions that are not always reliable over time, and that people's judgments can be swayed by details beyond the physical evidence being examined (e.g., the emotional context of case information) in cases that are ambiguous (e.g., an impression that is distorted, degraded, or highly similar to a nonmatching candidate impression; but see Hall & Player, 2008, and Schiffer & Champod, 2007, for studies finding no effects).

Separating Accuracy and Response Bias

There has been some published criticism of the methodology used in previous work on contextual influences (e.g., Saks, 2009, on Hall & Player, 2008). Specifically, the issues raised concern the use of performance measures that allow inconclusive judgments to be made, and the associated difficulty in capturing legitimate differences in discrimination ability. There has also been no work to date directly examining the influence of case information on examiners' performance accuracy (i.e., the ability to distinguish between print pairs that "match" from those that do not) versus their response bias (i.e., the extent to which participants say "match" or say "no match" regardless of the correct response).

Performance in previous studies has been measured by comparing the average number of details or "minutiae" in the fingerprint regarded as important by novice (e.g., Schiffer & Champod, 2007) and expert examiners (e.g., Langenburg, Champod, & Wertheim, 2009), comparing the mean percentage of total "match" responses made by novices (irrespective of whether the "match" decision was correct or not; e.g., Dror et al., 2005), or measuring intraexaminer reliability (i.e., the consistency of an examiner's judgments on the same case at different times; Dror et al., 2011; Dror & Charlton, 2006; Ulery, Hicklin, Buscaglia, & Roberts, 2012). Measures such as the total number or percentage of "match" responses, however, only tell a part of the story (i.e., frequency of correct identifications and false alarms), and fail to take into account the other half of possible performance outcomes: misses, and correct exclusions.

Here, we build on previous work by examining the impact of contextual information on people's forensic comparison decisions using separate measures of accuracy and response bias. By distinguishing between these two performance indicators, we can see precisely how—and by how much—contextual information influences human performance (see Thompson, Tangen, & McCarthy, 2013, for further discussion).

Fidelity, Generalizability, and Control

Besides measurement, another challenge in designing experiments on context effects is balancing *fidelity* (i.e., the degree of similarity between experimental conditions and the reference domain; Brunswik, 1956; Rasmussen, Pejtersen, & Goodstein, 1994; Thompson et al., 2013), *generalizability* (i.e., the extent to which the results are theoretically applicable to situations beyond those examined in the study; Thompson et al., 2013), and *control* (i.e., the extent to which experimenters are able to isolate and manipulate variables to detect genuine differences). The ideal experiment would have all three of these design characteristics, but often, one comes at the cost of the other (Sanderson & Grundgeiger, 2015; Thompson et al., 2013).

It is tempting to think that the gold standard would be an experiment that tests expert examiners (unbeknownst to them), and one that perfectly recreates specific work situations (e.g., covertly introducing contextual information into examiners' workflow). Indeed, there are some questions that can only be answered with this arrangement (e.g., gauging the performance of individual examiners). However, these high fidelity conditions come at a cost of reduced generalizability and reduced control. There is a great deal of variation in work conditions and practices across forensic laboratories, for example, making it difficult to apply the results of

high fidelity experiments to other laboratories (with different tools, workloads, workflows, etc.) or to the domain in general. Control wanes as well in these sorts of experiments as isolating variables and measuring performance can be difficult if examiners still have access to all their usual networks and tools (e.g., allowing inconclusive judgments; Saks, 2009; see also Thompson et al., 2013, on *Separate Accuracy and Response Bias*).

Likewise, experiments that opt for high control and generalizability, often suffer reduced fidelity—they are (by design) artificial and less like "real life" (Mook, 1983). Experiments of this nature are also important as they help to answer different questions (e.g., are context effects a result of changes in accuracy or the decision strategy employed?). The ultimate goal is to strike a balance between fidelity, generalizability, and control that best addresses the research question (Brinberg & McGrath, 1985; Mook, 1983). Eventually, with a large enough bank of studies, we can begin to examine the patterns that emerge from the converging evidence.

Our goal in the present series of experiments is to get an idea of whether the accuracy of people (in general) on a forensic comparison task can be influenced by case information and the prior experience of similar cases. In designing the experiments, we did not set out to imitate the day to day operations of a fingerprint unit. Instead, our goal was to achieve a high degree of control in manipulating the saliency and familiarity of case information and measuring their effect on both accuracy and response bias.

Through the Lens of Prior Experience

Forensic examiners, like the rest of us, tend to be attracted to a sense of *naïve realism*, believing that our raw perceptions are accurate and unbiased reflections of the world, uncontaminated by our preferences, preconceptions, prior experiences, and interpretations (Segall, Campbell, & Herskovits, 1966). Most of us also believe that human perception and memory work like a video camera, where we perceive the world through our senses, as a literal representation, and that the world always appears the same way to everyone. As plausible and inescapable as this "video camera" perspective might seem, it has some serious problems.

People experience the same objects and events very differently depending on our sensory organs (e.g., as many as one in 12 men are red/green color blind and will confuse blue and purple; Kaiser & Boynton, 1996), the context (e.g., the misleading information paradigm, Loftus, Miller, & Burns, 1978; the Deese-Roediger-McDermott task, Roediger & McDermott, 1995), and the experiences we have accumulated (e.g., our experiences with a top-lit world, three-dimensional shapes, light, and shading give rise to many compelling visual illusions; see Adelson, 1995; Shepard, 1990, 1992, and Thomas, Nardini, & Mareschal, 2010, for some examples).

Naïve realism is also the basis for *bias blindness* or the *not me fallacy* (Pronin, Lin, & Ross, 2002). When we are not aware of having made an interpretation we are blind to the fact that our judgments and decisions are easily swayed by the information available to us and by our prior experiences. The problem of bias blindness is nicely illustrated by the Chair of the Fingerprint Society in the United Kingdom, Martin Leadbetter. He provided the following response to findings by Itiel Dror and colleagues (e.g., Dror et al., 2005; Dror et al., 2006; Dror & Charlton, 2006) that contextual information (e.g., "the suspect confessed to the crime" or emotion-evoking case information) can sway the judgments made by experienced fingerprint examiners (Leadbetter, 2007):

> Any fingerprint examiner who comes to a decision on identification and is swayed either way in that decision making process under the influence of stories and gory images is either totally incapable of performing the noble tasks expected of him/her or is so immature he/she should seek employment at Disneyland.

In this case, Leadbetter fails to realize that the influence of contextual information is not deliberate and cannot be controlled (Nisbett & Wilson, 1977) any more than you can will away the effects of a visual illusion. Heuristics and cognitive biases are adaptive strategies that are based on our memory for prior instances, and allow us to arrive at rational conclusions, most of the time (Kahneman, 2011; Tversky & Kahneman, 1971, 1973, 1974). Forensic experts, like the rest of us, cannot simply will away their previous experiences and expectations, nor is this necessarily desirable (Tangen, 2013). We draw on this instance-based view of bias to investigate contextual influences across three experiments.

The Experiments

In the following series of experiments, we test groups of novices on their ability to discriminate between matching and nonmatching fingerprint pairs, measuring response bias and accuracy. In Experiment 1, novices are provided with case information rated as severe (vs. not severe) and are then asked to compare pairs of fingerprints. Our goal in this first experiment was to simply gauge whether case information, previously demonstrated to influence a person to say "match" at a higher rate (e.g., Dror et al., 2005) could also sway her accuracy. In Experiments 2 and 3, we go on to investigate one aspect of context that has been shown to influence accuracy in other domains of expertise (e.g., diagnostic medicine)—the familiarity of a case (Graber, Franklin, & Gordon, 2005; Norman, Young, & Brooks, 2007; Young, Brooks, & Norman, 2007). We test, first, how the familiarity of case information influences response bias and accuracy (Experiment 2), and then, in Experiment 3, go on to test how the familiarity of the target stimuli—the fingerprints themselves—influence response bias and accuracy after all other sources of case information have been removed.

Experiment 1

In Experiment 1, we manipulate crime "severity" (similar to Dror et al., 2005, which was referred to as "emotional context" in previous work) by presenting novices with pairs of fingerprints alongside case reports and images that are rated as either "severe" or "not severe" and measure their performance on a fingerprint comparison task. The question we wish to address in Experiment 1 is not about crime severity per se, but about measurement: whether a contextual factor previously demonstrated to sway people's decisions, such as the severity of the crime, will influence participants' response bias (i.e., their tendency to say "match") or whether it will affect their overall accuracy (i.e., sway an otherwise correct judgment to be incorrect or vice versa). Previous work has shown that novices tend to overcall matches in situations where the base rates for matching and nonmatching prints are 50/50 (e.g.,

Tangen et al., 2011). Given the same base rates in this experiment, if the crime severity has an effect on accuracy, then we might expect that participants will be less accurate on trials in which they are presented with case information and images of crimes that are more severe (compared to less severe). If, however, the results of Dror et al. (2005) were due to a shift in participants' response bias, we might expect that participants will simply respond more liberally (i.e., tend to say "match" regardless of whether the prints actually match or not) in the severe condition than in the less severe condition.

Method

Participants. Participants were 48 undergraduate psychology students from The University of Queensland participating in exchange for course credit. There were 32 females and 16 males with a mean age of 23 years. We used novice participants in each of our experiments to control for any prior experience with fingerprints as well as familiarity with the case information.

Design and performance measures. We employed a within-subjects design to manipulate the severity of the contextual information across two conditions (cases that have been rated as "severe" vs. cases rated as "not severe"). In order to measure response bias and accuracy, we used a forced choice confidence scale ranging from 1 (*sure different*) to 12 (*sure same*); ratings of 1 through 6 were counted as a "no match" response and ratings of 7 through 12 as a "match" (see Figures 1 and 2). Inconclusive judgments were not permitted using this design, allowing us to distinguish between accuracy and response bias (Green & Swets, 1996; for a more comprehensive breakdown of signal detection as a method used to measure performance, see Phillips, Saks, & Peterson, 2001, and Thompson et al., 2013).

Fingerprints. The fingerprints were the same as those used by Tangen et al. (2011) and sourced from the Forensic Informatics Biometric Repository. Tangen et al. (2011) lifted the crime scene or "latent" fingerprints (left by undergraduate students participating for course credit) from multiple surfaces (i.e., plastic, glass, wood, metal), documenting the source of each latent print (e.g., the person who deposited the print) to ensure that the ground truth was known. Matching prints were created by collecting fully rolled fingerprint exemplars from the same participants who deposited the latent fingerprints on a separate occasion. Highly similar but nonmatching pairs were created by Tangen et al. (2011) by entering each latent fingerprint into the Queensland Police Service fingerprint database, and using the most highly ranked nonmatching exemplar from the search. Overall, the set consisted of a total 36 fingerprint trios: a latent print, a corresponding matching print, and a highly similar but nonmatching print. Each of the 36 latent prints were randomly paired with the corresponding match or the corresponding nonmatch and each participant received 18 matching and 18 nonmatching pairs in a different random order. In each set of 18, nine pairs were accompanied by a severe case report and nine were accompanied by a less severe case report (selected at random). The experiment was, therefore, a 2 (Match, Nonmatch) × 2 (Severe, Less Severe) within-subjects design, where

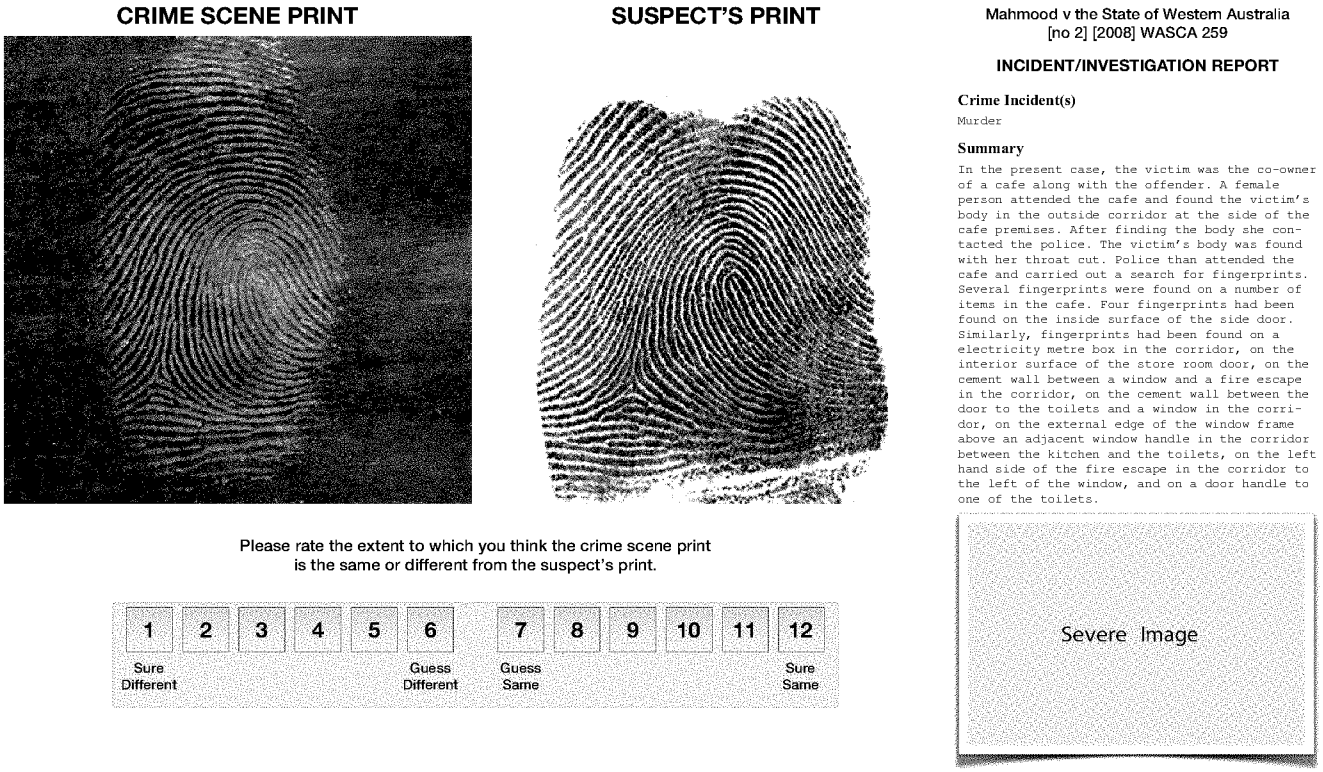

Figure 1. A screenshot from Experiments 1 and 2 of a pair of nonmatching fingerprints alongside a severe case report.

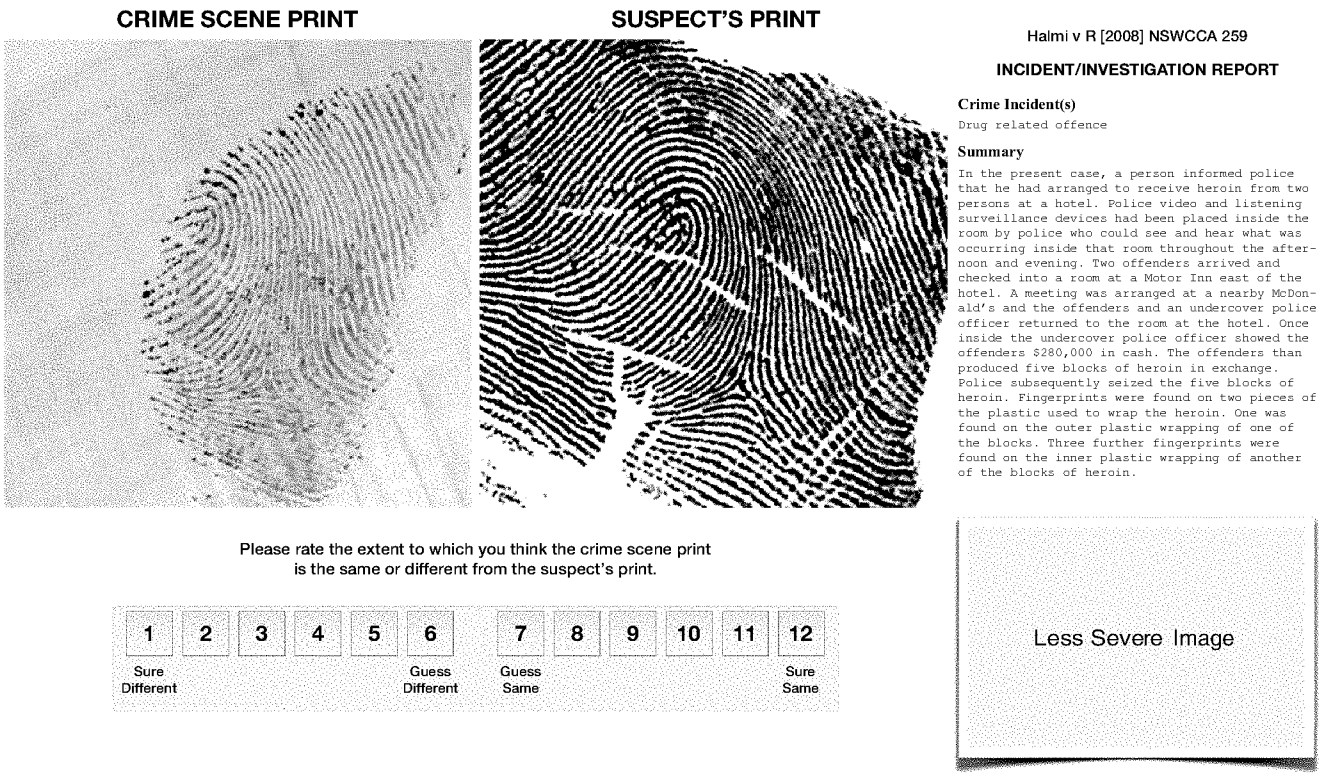

Figure 2. A screenshot from Experiments 1 and 2 of a pair of matching fingerprints alongside a less severe case report.

we combine the hit and false alarm rates into separate measures of discriminability and response bias. A pair of nonmatching fingerprints is depicted in Figure 1 alongside a severe case report and a pair of matching prints is depicted in Figure 2 alongside a less severe case report.

Incident reports and photographic stimuli. The case reports were sourced from case law databases such as LexisNexis and CaseBase. Thirty-six recent criminal cases from across Australia involving fingerprint evidence were selected. Eighteen of these cases related to crimes of murder, aggravated sexual assault, terrorism, assault, and armed robbery, and they were classified as "severe" (on the basis that they involved direct-physical harm to others). The remaining 18 cases related to crimes such as break and enter, drug related offenses, and theft and were classified as "less severe." Cases were summarized and presented as an incident/investigation report in a single paragraph format as depicted in Figures 1 and 2.

The photographs were sourced from the Google Images database. Each image was carefully selected to closely reflect the specific details of each of the 36 individual cases and depict a high level of realism. The images in the severe condition contained graphic visual material similar to Dror et al. (2005), and the less severe cases were also presented along with images (both sets are available by contacting the authors). The images selected for the severe condition were selected to resemble the injuries that might be received by a victim in the corresponding case. The images chosen in the less severe condition typically depicted photographs of items related to the corresponding crime (e.g., drugs, money, police dusting for fingerprints at a break and enter scene).

Pilot: Manipulation check for case severity. To confirm that the case reports and images accurately reflected novices' perceptions of case severity, we tested a separate group of 13 novices in a pilot study. Participants were presented with each of the incident/investigation reports and related photographs as detailed above alongside an image of a latent fingerprint in random order. Participants were instructed to rate the severity of each case on a scale from 1 (*not severe at all*) to 9 (*very severe*). As anticipated, participants rated the severe cases as more severe ($M = 7.04$) than the less severe cases ($M = 2.90$), $t(12) = 11.99$, $p < .001$, $d_{av} = 4.12$, 95% confidence interval (CI) [3.90, 4.47]. These pilot results indicate that our classification of severe and less severe cases were in line with participants' perceptions of case severity.

Procedure. After reading an information sheet about the experiment, participants were instructed to read the incident/investigation reports on the computer screen. To ensure that they were reading each case carefully, they were told that they would be asked about details of the cases later on in the experiment (i.e., we asked them how many cases involved weapons at the end of the experiment). Participants were also instructed to imagine they were expert fingerprint examiners who had the task of deciding whether the latent print found at a crime scene and a fully rolled suspect print were from the same person (see Appendix A for a complete set of instructions). They were then presented with the incident/investigation report and related image before completing

the fingerprint-matching task for all 36 cases as described above. To further ensure that participants read the investigation reports, the fingerprints were masked by a semitransparent gray mask until the participant indicated they had read the passage and were ready to compare the prints.

Results

To derive scores of response bias and accuracy separately, hit and false alarm rates were calculated for all participants in each condition. For example, confidence ratings of 7 or more (i.e., "match" responses) were coded as hits for the match trials and false alarms for the nonmatch trials (the raw confidence ratings for the three experiments are available by contacting the first author). Participants correctly declared matching fingerprints as a "match" for 83% of the severe cases, compared to 80% for the less severe cases. For nonmatching fingerprints, participants correctly declared them as a "nonmatch" for 45% of the severe cases, and 50% of the time for less severe cases. This pattern of results is similar to novice performers in Tangen et al. (2011) who were 75% correct for matching pairs and 45% correct for nonmatching pairs.

Participants' mean discrimination index (A'), or performance accuracy, was derived from their hit and false alarm rates in each condition (see Vokey et al., 2009, for a similar analysis and discussion). A' is a nonparametric measure that reflects the proportion of hits relative to false alarms, where an A' of 1 indicates perfect discrimination and an A' of 0.5 indicates chance discrimination (Donaldson, 1992). Participants' ability to discriminate between prints did not differ statistically between severe cases (*Mean* $A' = .72$) and less severe cases (*Mean* $A' = .70$), $t(47) = .77$, $p = .446$, $d_{av} = .14$, 95% CI [0.10, 0.17].

To assess response bias, B''_D was derived for each participant in each condition. A B''_D score of -1 reflects a strong liberal response bias (i.e., a tendency to say "match" more), a score of 1 reflects a strong conservative response bias (i.e., a tendency to say "no match" more), and a B''_D score of 0 reflects no bias (Donaldson, 1992). We found that participants had a strong liberal response bias and tended to say "match" overall, which is consistent with novice data in previous studies (e.g., Tangen et al., 2011). Moreover, as has been found previously, participants were biased to say "match" more often when the prints were accompanied by severe case information (*Mean* $B''_D = -.56$; an average of 12.42 out of 18 severe cases were rated as a match), compared to the less severe case information (*Mean* $B''_D = -.44$; an average of 11.71 out of 18 less severe cases were rated as a match), which was demonstrated using a two-tailed paired t test revealing a significant difference between these conditions, $t(47) = 2.05$, $p = .046$, $d_{av} = .24$, 95% CI [0.10, 0.38].

Cross-experiment comparison in signal detection space. We then plotted our results in a signal detection diagram (see Figure 3) in order to compare performance on the severe versus less severe trials and to compare results across experiments (see Thompson et al., 2013). This diagram is an illustration of participants' mean performance for the severe versus less severe trials, plotted in signal detection space (i.e., the space of all possible responses): where discrimination accuracy is represented by the vertical axis (the top indicates perfect discrimination and the bottom chance discrimination) and response bias is represented along the horizontal axis (the far left of the diagram indicates a tendency to say "match" on all trials, the far right indicates a tendency to say "no match" on all trials, and the middle of the axis indicates a response bias that perfectly reflects actual base rates of matching and nonmatching trials), which is 50/50 here. The closer the data points are to the top of the diagram, the more accurate participants performed in that condition (i.e., the data points for both conditions are in roughly similar positions along the vertical axis, reflecting the similar discrimination scores for severe and less severe trials). Their position on the left or right of the diagram indicates their response bias (e.g., the data point for the severe trials is further to the left than the less severe trials, reflecting participants' more liberal response in this condition). The contingency scores used to plot our results were derived by computing the mean hits, false alarms, correct rejections and misses for both conditions, and by converting these to percentages (see Thompson et al., 2013). We refer to this diagram again in Experiments 2 and 3 to provide a broader context to each set of results.

Discussion

When participants in Experiment 1 were asked to compare pairs of fingerprints, which were presented alongside severe or less severe case reports, their ability to discriminate between the matching and nonmatching prints was unaffected by the graphic nature of the report. Their response bias, or their tendency to say "match," however, was affected. That is, participants were slightly more likely to say "match" (a liberal response bias) when presented with the severe case information.

Our present design did not permit us to examine the cognitive or motivational factors responsible for the effect. However, the results from Experiment 1 demonstrate that in order to gauge whether case information increases or decreases people's accuracy, measures of performance need to account for both ways of being right (i.e., hits and correct rejections) and both ways of being wrong (i.e., false alarms and misses). While perceived case severity may sway people to say "match" more, it is misleading to conclude from these results that removing this case information will reduce the likelihood of error across the board. More accurately, removing information about the type of crime may reduce the number of false identification errors in severe cases, but in doing so, the amount of miss errors may increase. As a result, we should be cautious about drawing conclusions about the influence of case information on error rates from studies without separate measures of response bias and discrimination ability.

Experiment 2

In Experiments 2 and 3, we build on the findings of Experiment 1 by investigating a source of contextual information that we predict will have a significant impact on accuracy—namely, the familiarity of the information. Previous research in diagnostic medicine has shown that familiar nondiagnostic information (e.g., patient demographic details similar to previously encountered cases) can sway the judgments of novice diagnosticians. That is, more weight is given to diagnoses that are cued by the familiar case information (compared to an equally plausible alternative diagnosis; Young, Brooks, & Norman, 2011). This research suggests that diagnosticians store information about previous cases in memory and use these memories to aid current decision-making (Kelley & Jacoby, 1996; Young et al., 2011).

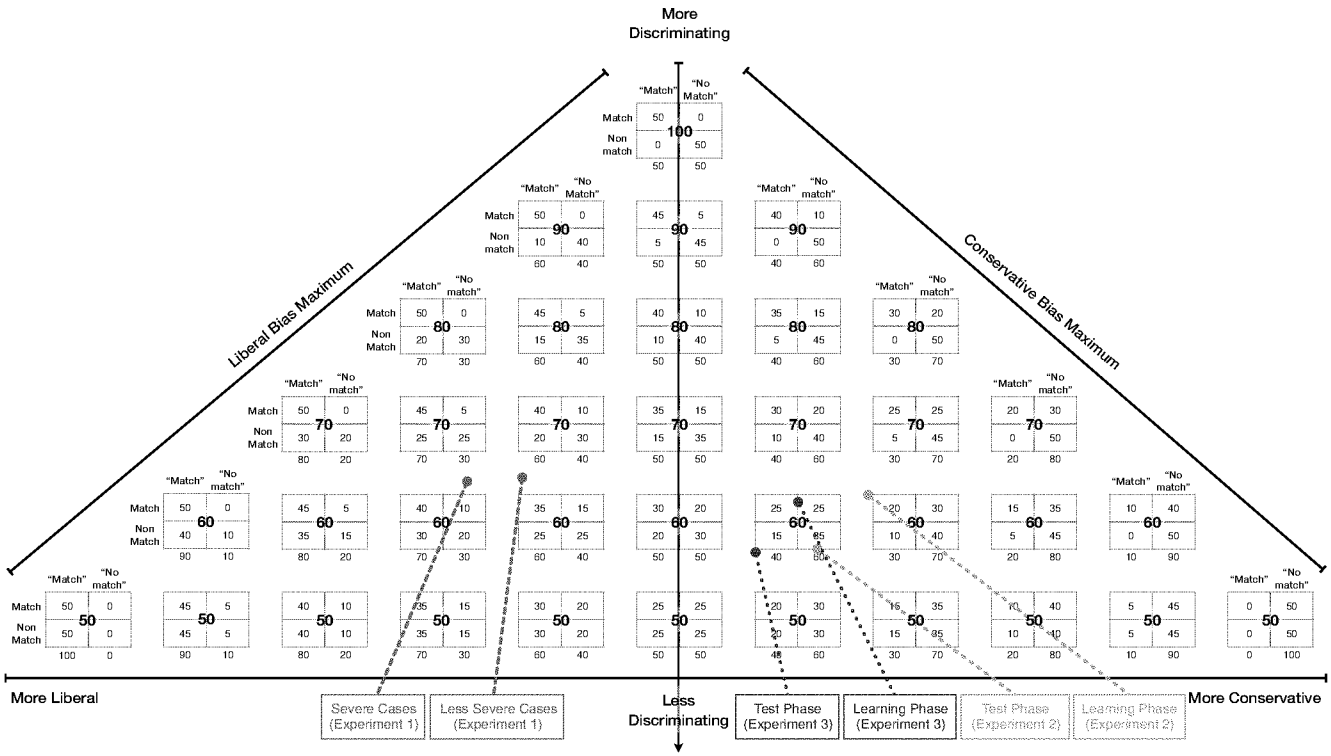

Figure 3. The space represents all possible performance results from a discrimination task. Each of the tables that comprise the figure is a 2 × 2 contingency table depicting the four possible outcomes where two prints match or not and someone labels them a "match" or "no match." The numbers align with hits, false alarms, misses, and correct rejections. The large number in bold at the center of each table depicts the sum of the two diagonal cells ranging from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. The column totals at the bottom of each table depict response bias with a liberal bias (a tendency to say "match") depicted on the left of the figure and a conservative bias (a tendency to say "no match") depicted on the right of the figure. The data points in the space are the locations of the actual results for each of the three experiments, where each filled circle represents the center of the 2 × 2 contingency table based on the data from each. See the online article for the color version of this figure.

Like diagnosticians, forensic examiners are often exposed to a range of materials that are related to the case at hand when making an identification. Given the similarity between the classification tasks performed in diagnostic medicine and the comparison tasks performed by fingerprint examiners, it is possible that the accuracy of fingerprint examiners' decisions may also be influenced by the familiarity of the case information. Can knowing the details about a case influence someone's decision about a completely different case if the two cases are similar? Specifically, can the link between the details of a case (e.g., events leading up to the crime, the nature of an injury, the location of the victim's home) and the outcome of the case (e.g., the crime scene print matched the suspect) influence someone's judgment about a completely different case that is similar, but with the opposite outcome (e.g., the crime scene print does not match the suspect)?

In Experiment 2, we test this claim by first presenting participants with a series of cases and fingerprint pairs as we did in Experiment 1. Half of the prints match and half do not, and participants are told whether they made the correct decision or not. In the second half of the experiment, we present a series of new cases that are very similar to those they have just seen. Nearly every detail about the case is altered slightly. For example, "The intruder then snatched an iPad out of the victim's hand and threw it to the floor" is changed to "The intruder then snatched a book out of the victim's hand and threw it to the other side of the bed" in the similar case. The fingerprints that are presented alongside the case report are completely new, but if the case was presented alongside a pair of matching prints in the first half, then the similar case was presented alongside a nonmatching pair in the second half. Similarly, if the case was presented with nonmatching prints in the first half, then the similar case was presented with matching prints in the second half. Participants were not told during the second phase whether they were correct or not, as we are interested in whether they change their decision from the first to the second half. If their accuracy drops significantly, then this is a clear indication that they are sensitive to the familiarity of the case. Specifically, we expect familiar case information to sway people's judgments in the direction of previous similar cases (e.g., if a previous case was a "match," then people should be more likely to rate the novel similar case as a "match" as well—even if it is not). In Experiment 2, we simply manipulate the *extrinsic* familiarity of cases (i.e., similarity of information in the case reports to previ-

ously encountered cases). We go on to manipulate the *intrinsic* familiarity of cases (e.g., similarity of the fingerprints in the case to previously encountered cases) in Experiment 3.

Method

Participants. Participants were 35 undergraduate psychology students from The University of Queensland who participated in exchange for course credit. There were 26 females and nine males with a mean age of 25 years.

Design and performance measures. Experiment 2 is split into two halves: a learning phase (18 trials) and a test phase (18 trials). In the first half—the learning phase—participants are presented with case reports and fingerprint pairs as in Experiment 1. Half of the fingerprints match, and half do not. Participants are told whether their decision was correct or not during the learning phase. In the second half—the test phase—participants are presented with the same 18 case reports as the learning phase (in a different random order for each participant), but each sentence in the report has been modified slightly so the two reports are very similar. None of the 18 fingerprint pairs in the test phase were presented in the learning phase and participants were not provided with feedback on their decision during the test phase. However, if the prints that accompanied a particular case in the learning phase matched, then the new set of prints that accompanied the similar case in the test phase did not match, and vice versa. If participants learn the association between the information in the report and the outcome (i.e., match or no match), and if they retrieve this association during the test phase when prompted with a very similar case, then we expect that they will say "no match" to the matching prints and "match" to the nonmatching prints during the test phase. That is, if participants are sensitive to the similarity between the two cases, then we should see a significant decrease in their accuracy during the test phase. The experiment was, therefore, a 2 (Phase 1, Phase 2) × 2 (Match, No Match) within-subjects design, where we combine the hit and false alarm rates into separate measures of discriminability and response bias as we did in Experiment 1.

Fingerprints. The fingerprints used in Experiment 2 were the same 36 sets used in Experiment 1. They were counterbalanced across phase by randomly selecting 18 pairs (nine matching and nine nonmatching) for use in the learning phase. The remaining 18 sets (nine matching and nine nonmatching) were used in the test phase.

Incident/investigation reports. There were 36 written investigation reports presented in random order to each person. Half of the reports were the same as those in Experiment 1, with nine from the pool of severe cases and nine from the pool of less severe cases. These 18 cases were used in the learning phase as they were written in Experiment 1. Similar to Young et al. (2011), the other 18 cases in the test phase were carefully designed to be highly similar to the learning cases, but not identical. We achieved this by slightly altering each aspect of each case report for the test trials (see Appendix B for a complete example of the similarity manipulation).

Procedure. Participants were given the same instructions as in Experiment 1, which were presented on the computer screen (see Appendix C for a complete set of instructions). When participants clicked begin, they were presented with the learning phase, where they were presented with the 18 original written investigation reports described earlier and completed the fingerprint comparison task on all 18 trials. Similar to Young et al. (2011), in the learning phase, participants were provided with immediate feedback following each trial on their performance (i.e., presented with a response of either "Correct" or "Incorrect," plus correct/incorrect audio cues)—this was to ensure that all participants were provided with the same information about the correct responses during learning for the familiarity manipulation. Participants then immediately completed the test phase in which they were presented with the 18 highly similar investigation reports, and again completed the fingerprint comparison task for each trial. No feedback was provided during the test phase, again similar to Young et al., (2011). Participants were also asked to report the relevant case information that informed their decision as an added measure to ensure that they read the reports.

Results

We used the same method as Experiment 1 to derive A' and B''_D as measures of discrimination accuracy and response bias respectively, and plotted our results in Figure 3. As predicted, the mean discrimination in the learning phase (*Mean* A' = 0.7; 47.29% of match cases were rated as "match" and 76.43% of nonmatch cases were rated as a "no match") was greater than the test phase (*Mean* A' = .62; 46.43% of match cases were rated as "match" and 67.86% of nonmatch cases were rated as a "no match"). As can be seen in Figure 3, the data point for the test phase is lower on the vertical axis than the data point for the learning phase trials, reflecting the observed decrease in discriminability. A two-tailed paired t test confirmed this difference to be significant, $t(34) = 2.57$, $p = .015$, $d_{av} = .59$, 95% CI [0.55, 0.65].

In contrast to Experiment 1, participants had a conservative response bias and tended to say "no match" more than "match" in both phases of the experiment (see Figure 3, where the data points from Experiment 2 are located much further to the right than Experiment 1). Participants also adopted a more conservative response bias on the trials during the learning phase (*Mean* B''_D = .40; an average of 6.09 out of 18 learning phase cases were rated as a "match") compared to the test phase (*Mean* B''_D = .25; an average of 6.86 out of 18 test phase cases were rated as a "match"), as illustrated in Figure 3 where the data point in the learning phase is located further to the left than the test phase. A two-tailed paired t test also revealed this difference to be significant, $t(34) = 2.65$, $p = .013$, $d_{av} = .24$, 95% CI [0.05, 0.47].

Discussion

The aim of Experiment 2 was to determine whether an association between the details and the outcome of the case can influence someone's judgment about a different case that is similar. Our results demonstrate that participants were sensitive to the similarity of the case information contained in the reports between the two phases of the experiment. That is, during the first phase, participants must have learned the associations between the information in the case reports and the outcome (i.e., match/no-match), as evidenced by those associations influencing outcome decisions during the test phase. While objectively irrelevant, the presentation of similar information altered participants' decisions about novel pairs of prints. Participants tended to say "no-match" to the match-

ing pairs and "match" to the nonmatching pairs, resulting in a significant drop in accuracy. They also demonstrated a significant shift in response bias between the two phases, where they tended to say "match" less often during the learning phase compared to the test phase.

These results indicate that participants made fingerprint comparison decisions in line with the correct response that was associated with the similar previously encountered case. These findings are similar to observations in the diagnostic medicine literature, which support the idea that people use the knowledge gained from previous experiences (previous similar case reports in this case) to aid in current decision-making on a forensic comparison task. To further understand these findings, we explore two possible explanations for the results.

Instance-based retrieval hypothesis. One explanation is that the similarity of each case report in the test phase acted as a cue for the rapid recall of the similar prior instances in the learning phase, creating a "feeling of knowing" or sense of fluency and familiarity (Kelley & Jacoby, 1996; Young et al., 2011). Unlike the experience of recollection (an analytic or explicit recognition process), the experience of familiarity (a nonanalytic or implicit recognition process) cannot be pinpointed to its exact source and thus affords less control over the influence of a particular prior experience (Kelley & Jacoby, 1996). It is this lack of control that may have increased people's reliance on extrinsic but familiar cues when comparing the fingerprints. Our results provide evidence that at least some information from the cases encountered in the learning phase is implicitly (or perhaps explicitly) leaking over to participants' decisions in the test phase.

Feedback hypothesis. In order to ensure that each participant was aware of the correct response for later retrieval on familiar trials, it was necessary to provide them with feedback during the learning phase just as Young et al. (2011) did in their experiment. Presenting feedback, however, introduces another possible explanation for the decrease in performance during the test phase in that removing feedback during the second half of the experiment may account for the reduction in performance (as well as the more liberal response bias). Recent experiments on the effects of feedback on fingerprint comparison decisions (Searston & Tangen, in preparation) have consistently shown improvements—not decrements—in performance where feedback was presented during practice but removed during a test of transfer. If the same was true in Experiment 2, then it would suggest that the effect of similarity was even greater than we observed because this effect of feedback during practice would have counteracted the influence of similarity in this design.

The finding that the learning benefits of feedback remain steady once trial-by-trial feedback is removed, is quite robust—having been demonstrated in learning studies across several domains (e.g., Wulf & Schmidt, 1989; see also Salmoni, Schmidt, & Walter, 1984, for a review of similar effects in motor learning). Indeed, experiments in a very similar visual discrimination domain—unfamiliar face matching—have demonstrated that the benefit of feedback remains after training, even when participants are tested on a completely different, more variable, set of images (e.g., White, Kemp, Jenkins, & Burton, 2014). Even visual discrimination studies involving significantly more trials than Experiment 2 (e.g., 12 blocks of 80 trials vs. our 36 trials), where factors such as fatigue would be more likely, have shown that removing feedback after as many as eight blocks of training does not result in a deterioration in performance (e.g., Herzog & Fahle, 1997). If tested immediately, as in this experiment, it is likely that participants still have access to the mental representations developed during the initial phase, resulting in the stickiness of the feedback effect observed in the literature.

This phenomenon is also consistent with current theories of learning, including the work around "desirable difficulties" showing that retrieval practice or testing, similar to our test phase, can be a powerful learning event in and of itself—even when corrective feedback is not provided (Bjork, 1975; Bjork & Bjork, 2011; Landauer & Bjork, 1978). Participants' accuracy over the last 18 trials in Experiment 1 (*Mean A'* = 0.73) did not decline from the first 18 trials (*Mean A'* = 0.71) suggesting that fatigue was not decreasing performance over the same number of trials in Experiment 1. Taken together, this previous body of work, and the results from Experiment 1, suggest that fatigue, or the removal of feedback, are not convincing explanations for the observed decrement in accuracy in Experiment 2.

Accounting for response bias effects. We suspect that the feedback we provided might account for the conservative response bias that participants adopted in Experiment 2 (compared to the liberal bias of those in Experiment 1). That is, participants are likely unaware of how highly similar a pair of fingerprints can be from two different people until they see examples of these materials with feedback. This suspicion is supported by Thompson et al. (2013) who demonstrated that trainee fingerprint examiners tend to become more conservative with training, which presumably involves experience with highly similar exemplars and feedback from senior examiners.

The presence of feedback during learning might also have resulted in participants overcorrecting more during learning than on the test, where no feedback was provided. Another possibility for participants saying "match" more often for familiar cases is that the matching pairs encountered during learning were more memorable than the nonmatching pairs (Kelley & Jacoby, 1996). In other words, this finding could be related to the confirmation bias (i.e., the tendency to search for and interpret information in terms of positive rather than negative instances) such that a "match" is more likely to be remembered (and more likely to influence current decision making) because it is a positive event. In other words, the improved memory for matching relative to nonmatching pairs might be equivalent to the belief, held by many, that arthritis pain is influenced by the weather because those individuals notice their pain more during an extreme weather event (a positive event) but pay less attention when the weather is fine (Redelmeier & Tversky, 1996).

Experiment 3

In Experiments 1 and 2, we examined two ways that *extrinsic* case information (e.g., case reports that provide a description of the crime) could influence people's performance on a forensic comparison task. In Experiment 3, we examine the influence of *intrinsic* familiarity by testing whether familiar fingerprints can sway people's decision making—a source of information that cannot be removed from examiners' workflow. Specifically, we replicate the procedure from Experiment 2, but we replace similar *case information* with similar *fingerprint pairs* and examine whether people

are swayed by this familiar information. If people are blinded to all extrinsic case information, can their performance still be swayed by the similarity of the fingerprints to previously encountered fingerprints?

Previous research in diagnostic medicine suggests that people may indeed be swayed by intrinsic familiarity. For example, the presence of diagnosis-relevant information that is similar to specific prior experiences can strongly influence diagnostic reasoning in both doctors and students (Allen, Norman, & Brooks, 1992; Brooks, Norman & Allen, 1991; Hatala, Norman, & Brooks, 1999). In these experiments, the similarity of visual stimuli (e.g., skin lesions) was manipulated and the overall similarity of stimuli to previously encountered cases was found to influence clinical reasoning among novices and experts. Participants in these studies assigned more weight to a familiar symptom description, and they were more likely to diagnose a fictional patient with the diagnosis supported by a symptom description that they previously encountered. Research in other medical fields has further demonstrated this effect with written case materials suggesting that a reliance on past experiences in clinical reasoning is not limited to visual stimuli (e.g., Young et al., 2007). On the basis of this prior research, we predict that the visual similarity of the fingerprints to previously encountered cases will influence people's fingerprint comparison decisions in a similar fashion.

Method

Participants. Participants were 38 undergraduate psychology students from The University of Queensland, participating for course credit. There were 24 females and 14 males with a mean age of 27 years.

Design and performance measures. Experiment 3 employed the same within-subjects design, measures, methodology, and fingerprints as Experiment 2, except that participants were not provided with case reports about the crimes or any other information about the cases. We presented the full 36 pairs of fingerprints in a learning phase, and presented the 36 latent prints again during the test phase, but with the opposite outcome. That is, if the latent prints were paired with a matching print during the first half of the experiment, they were paired with a nonmatching print in the second half, and if they were paired with a nonmatching print during the first half, they were paired with a matching print in the second half. This methodology ensured that the latent prints remained the same during the learning and test phase, but that the comparison prints differed between the conditions—creating pairs of prints that were similar, but not identical. The experiment was, therefore, a 2 (Phase 1, Phase 2) × 2 (Match, No Match) within-subjects design, where we combine the hit and false alarm rates into separate measures of discriminability and response bias.

Procedure. In Experiment 3, participants were simply instructed to imagine they were expert fingerprint examiners who had the task of deciding whether the latent print found at a crime scene and the fully rolled suspect print were from the same person (see Appendix D for a complete set of instructions). They compared the 36 novel pairs of prints during the learning phase (i.e., a random presentation of 18 matching and 18 nonmatching pairs) and were provided with feedback on their decisions before comparing the 36 "familiar" pairs of prints during the test phase in random order without feedback.

Results

We calculated A' and B''_D using the same method as Experiments 1 and 2 and plotted our results in the signal detection diagram in Figure 3. Consistent with findings of Experiment 2, the mean discrimination in the learning phase (*Mean* A' = 0.69; 50.62% of match cases were rated as "match" and 71.20% of nonmatch cases were rated as a "no match") was greater than the test phase (*Mean* A' = 0.62; 52.91% of match cases were rated as "match" and 63.99% of nonmatch cases were rated as a "no match"), $t(37) = 2.58$, $p = .014$, $d_{av} = .54$, 95% CI [0.50, 0.59]. Participants also demonstrated a more conservative response bias (i.e., they tended to say "no match" more than "match") in the learning phase (*Mean* B''_D = .29; an average of 14.03 out of 36 learning phase cases were rated as a "match") compared to the test phase (*Mean* B''_D = .16; an average of 15.36 out of 36 test phase cases were rated as a match) of the experiment. A two-tailed paired t test revealed this difference to be significant as well, $t(37) = 2.13$, $p = .039$, $d_{av} = .22$, 95% CI [0.03, 0.42].

Discussion

Just as in Experiment 2, participants were clearly sensitive to the similarity of the fingerprint pairs as reflected in the drop in accuracy between the two phases of the experiment. This result is consistent with previous studies in diagnostic medicine that demonstrate a similar effect regarding the familiarity of medical imaging stimuli (e.g., Allen et al., 1992; Brooks et al., 1991).

Given that the latent fingerprints were the same in the learning and test phases of the experiment, it is possible that participants were relying somewhat on explicit recollection of the exact prior instances of the latent prints, as opposed to relying solely on an implicit feeling of familiarity. Theories of recognition memory, however, suggest that it may be easier to recall specific details when experiencing recollection, resulting in more control over the influence of that prior experience (Kelley & Jacoby, 1996). If participants were relying more on explicit recollection, recalling details about the previous pair, then we might expect them to be less easily swayed by their prior experience on the test trials, which would dampen the resulting effect of decreased accuracy at test. Participants in Experiment 2 and 3 still showed a significant decrease in accuracy from the learning phase to the test phase, suggesting that a reliance on the familiarity of the cases is more likely to be responsible for the effect. It also seems unlikely that participants were able to remember specific details of the prints across 72 trials, particularly given previous demonstrations of their poor explicit memory for similar fingerprint pairs (Thompson & Tangen, 2014).

In any case, whether participants are relying on a feeling of familiarity or explicit recognition of previous cases, the results of Experiments 2 and 3 demonstrate how something as subtle as the similarity of a case—even the similarity of a fingerprint pair—to a previous encounter can have a marked influence on current decision-making.

Implications

The role of familiarity in forensic comparison decisions is well worth examining further. As computerized databases of finger-

prints grow, the chance of finding a highly similar print from different individuals must necessarily increase (Dror & Mnookin, 2010). The same applies to examiners' experience; the more experience they gain with instances of fingerprint pairs, the more likely it is that they will encounter novel fingerprint pairs that are highly similar but not identical to previously encountered fingerprint pairs. Unlike extrinsic case information, examiners cannot be blinded to the familiarity of a fingerprint. Simply removing all extrinsic case information will not necessarily result in judgments that are completely objective or free from bias, nor is this a bad thing in every case (e.g., cases where the similar prior experience is consistent).

It remains to be seen if the influence of similar prior cases grows or decays with experience. Perhaps after seeing thousands of cases, they all begin to blend together resulting in a smaller effect. Alternatively, drawing on many similar prior cases could result in a larger effect. It is important to note that even though similarity decreased accuracy in Experiments 2 and 3, this drop in performance is part of the experimental design that was necessary to test our hypotheses. In many natural situations, similarity is a valid cue that could improve decision-making. That is, without switching the correct response from learning to test, we would expect the familiarity of the cases to result in an increase in accuracy. Identifying when familiarity is likely to help and when it is likely to hurt could inform the design of workplace systems that lead to examiners making more correct judgments.

General Discussion

In the present paper, we have outlined some of the problems with the "Disneyland" perspective of bias in forensic science (e.g., Leadbetter, 2007). Forensic examiners, like the rest of us, tend to believe that their raw perceptions are accurate and unbiased reflections of the world, uncontaminated by their preferences, preconceptions, and interpretations (Segall et al., 1966). Several experiments have now demonstrated otherwise: contextual information can sway the judgments made by even the most diligent examiners (see Kassin et al., 2013 for review). After the threat of contextual bias featured heavily in the U.S. National Academy of Sciences Report on the state of forensic science (National Research Council, 2009), the literature on the topic has grown, and many forensic laboratories are beginning to introduce blinding procedures (Risinger et al., 2014).

Such demonstrations of contextual influences in forensic science are certainly a good start, but they do not go far enough. These studies have demonstrated that a person's judgments *can* be swayed by contextual information (e.g., Dror et al., 2005) or that examiners might not be consistent in their judgments from one time to the next (e.g., Dror et al., 2006; Dror & Charlton, 2006), but these demonstrations have relied on contextual information in the most obvious sense (i.e., information about the case or the trace evidence designed to explicitly sway examiners' judgments, such as a confession). Instead, we have offered an instance-based conception of contextual influences where seemingly irrelevant and subtle information can sway people's judgments if it is sufficiently similar to the case or trace at hand.

Across three experiments, we demonstrated that case information can sway people's judgments by shifting their response criterion, but this shift does not always reduce their accuracy. For example, case information that is perceived to be more severe led people to respond more liberally, but their accuracy remained unchanged (Experiment 1). The familiarity of the case, on the other hand, did affect their accuracy, in the direction of previous similar cases (i.e., if a previous similar case was a "match," then people were more likely to also rate a novel case as a "match"—even though it was not—as demonstrated in Experiments 2 and 3). Most interestingly, the influence of familiarity on performance remained in Experiment 3, even when we removed all extrinsic case information. Our results add to the previous literature on context effects by demonstrating that the context does not have to be explicit or obvious in any sense to significantly affect performance. Crime severity or surface similarity does not explicitly implicate a particular suspect or judgment and it is this subtlety that makes these context effects compelling.

The context effects observed in our experiments may be even greater for examiners bearing the weight of genuine casework decisions. Others may wish to examine whether conditions that more closely resemble actual casework would increase the strength of the effects that we have shown here. Very little is known about the factors that affect expert performance in other areas of forensic science (fire investigation, blood pattern analysis, firearm and tool mark comparison, shoe print examination etc.). Another line of research might investigate whether effects of familiarity generalize to areas dealing with different and more variable trace evidence.

The approach that we adopted in the current set of experiments was to strip back the situation and introduce one difference at a time in the information presented (e.g., severe vs. less severe, familiar vs. unfamiliar), under laboratory conditions. This allows us to move closer toward isolating potential mechanisms that drive contextual influences. We have also presented a novel methodological approach to measuring the impact of sources of contextual information in forensic science, which may be useful to other researchers investigating this issue across applied domains. We need more studies like those we present here and others on this topic to get a better handle on the role that contextual influences play in forensic decisions—across situations, people, and cases. By adopting a narrow view of the contamination or threat of cognitive bias, the well intentioned pursuit of controlling for bias may have the unintended effect of stifling legitimate discussion about how to harness human expertise and improve the system in which forensic examiners work (Institute of Medicine, 2000).

References

Adelson, E. H. (1995). *Checkershadow illusion*. Retrieved from http://persci.mit.edu/gallery/checkershadow

Allen, S. W., Norman, G. R., & Brooks, L. R. (1992). Experimental studies of learning dermatological diagnosis: The impact of examples. *Teaching and Learning in Medicine, 4,* 35–44. http://dx.doi.org/10.1080/10401339209539531

Bjork, E. L., & Bjork, R. A. (2011). *Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning.* Psychology and the real world: Essays illustrating fundamental contributions to society. New York, NY: Worth Publishers.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Brinberg, D., & McGrath, J. E. (1985). *Validity and the research process.* Beverly Hills, CA: Sage.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120,* 278–287. http://dx.doi.org/10.1037/0096-3445.120.3.278

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd revised edition). Berkeley, CA: University of California Press.

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.

Campbell, A. (2011). *The fingerprint inquiry report.* Retrieved from http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127-2.html

Champod, C. (2014). Research focused mainly on bias will paralyse forensic science. *Science & Justice, 54,* 107–109. http://dx.doi.org/10.1016/j.scijus.2014.02.004

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.

Cole, S. A. (2004). Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again. *The American Criminal Law Review, 41,* 1189–1208.

Cole, S. A. (2013). Implementing counter-measures against confirmation bias in forensic science. *Journal of Applied Research in Memory & Cognition, 2,* 61–62. http://dx.doi.org/10.1016/j.jarmac.2013.01.011

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121,* 275–277. http://dx.doi.org/10.1037/0096-3445.121.3.275

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a "target" comparison. *Forensic Science International, 208,* 10–17. http://dx.doi.org/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., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156,* 74–78. http://dx.doi.org/10.1016/j.forsciint.2005.10.017

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. http://dx.doi.org/10.1016/j.jarmac.2013.02.003

Dror, I. E., & Mnookin, J. L. (2010). The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law Probability and Risk, 9,* 47–67. http://dx.doi.org/10.1093/lpr/mgp031

Dror, I. E., Péron, 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. http://dx.doi.org/10.1002/acp.1130

Expert Working Group on Human Factors in Latent Print Analysis. (2012). *Latent print examination and human factors: Improving the practice through a systems approach.* Washington, DC: U.S. Government Printing Office.

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

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165,* 1493–1499. http://dx.doi.org/10.1001/archinte.165.13.1493

Green, D. M., & Swets, J. A. (1996). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Haber, R., & Haber, L. (2013). The culture of science: Bias and forensic evidence. *Journal of Applied Research in Memory & Cognition, 2,* 65–67. http://dx.doi.org/10.1016/j.jarmac.2013.01.005

Hall, L. J., & Player, E. (2008). Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Science International, 181,* 36–39. http://dx.doi.org/10.1016/j.forsciint.2008.08.008

Hatala, R. J., Norman, G. R., & Brooks, L. R. (1999). Influence of a single example on subsequent electrocardiogram interpretation. *Teaching and Learning in Medicine, 11,* 110–117. http://dx.doi.org/10.1207/S15328015TL110210

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research, 37,* 2133–2141. http://dx.doi.org/10.1016/S0042-6989(97)00043-6

Heyer, R., & Semmler, C. (2013). Forensic confirmation bias: The case of facial image comparison. *Journal of Applied Research in Memory & Cognition, 2,* 68–70. http://dx.doi.org/10.1016/j.jarmac.2013.01.008

Institute of Medicine. (2000). *To err is human: Building a safer health system.* Washington, DC: National Academy Press.

Kahneman, D. (2011). *Thinking fast and slow* (1st ed.). New York, NY: Farrar, Straus & Giroux.

Kaiser, P. K., & Boynton, R. M. (1996). *Human color vision* (Vol. 287, pp. 353–500). Washington, DC: Optical Society of America.

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. http://dx.doi.org/10.1016/j.jarmac.2013.01.001

Kelley, C. M., & Jacoby, L. L. (1996). Memory attributions: Remembering, knowing and feeling of knowing. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 287–307). Mahwah, NJ: Erlbaum.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, United Kingdom: Academic Press.

Langenburg, G., Champod, C., & Wertheim, P. (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *Journal of Forensic Sciences, 54,* 571–582. http://dx.doi.org/10.1111/j.1556-4029.2009.01025.x

Leadbetter, M. (2007). Letter to the Ed. *Fingerprint World, 33,* 231.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31. http://dx.doi.org/10.1037/0278-7393.4.1.19

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387. http://dx.doi.org/10.1037/0003-066X.38.4.379

National Research Council, Committee on Identifying the Needs of the Forensic Science Community. (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: The National Academies Press.

Nisbett, R., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41,* 1140–1145.

Phillips, V. L., Saks, M. J., & Peterson, J. L. (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Sciences, 46,* 294–308.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369–381. http://dx.doi.org/10.1177/0146167202286008

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering.* New York, NY: Wiley.

Redelmeier, D. A., & Tversky, A. (1996). On the belief that arthritis pain is related to the weather. *PNAS Proceedings of the National Academy of*

Sciences of the United States of America, 93, 2895–2896. http://dx.doi.org/10.1073/pnas.93.7.2895

Risinger, D. M., Thompson, W. C., Jamieson, A., Koppl, R., Kornfield, I., Krane, D., . . . Zabell, S. L. (2014). Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science." *Science & Justice, 54,* 508–509. http://dx.doi.org/10.1016/j.scijus.2014.06.002

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. http://dx.doi.org/10.1037/0278-7393.21.4.803

Saks, M. J. (2009). Concerning L. J. Hall, E. Player, "Will the introduction of an emotional context affect fingerprint analysis and decision-making?" [Forensic Sci. Int. 181 (2008). 36–39]. *Forensic Science International, 191,* e19. http://dx.doi.org/10.1016/j.forsciint.2009.06.011

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95,* 355–386. http://dx.doi.org/10.1037/0033-2909.95.3.355

Sanderson, P., & Grundgeiger, T. (2015). How do interruptions affect clinician performance in healthcare? Negotiating fidelity, control, and potential generalizability in the search for answers. *International Journal of Human-Computer Studies, 79,* 85–96. http://dx.doi.org/10.1016/j.ijhcs.2014.11.003

Schiffer, B., & Champod, C. (2007). The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. *Forensic Science International, 167,* 116–120. http://dx.doi.org/10.1016/j.forsciint.2006.06.036

Searston, R. A., & Tangen, J. M. (in preparation). *Feedback, elaboration and contrast practice: three ways to boost visual category learning.*

Segall, M., Campbell, D., & Herskovits, M. J. (1966). *The influence of culture on visual perception.* New York, NY: The Bobbs-Merrill Company.

Shepard, R. N. (1990). *Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art.* New York, NY: WH Freeman/Times Books/Henry Holt & Co.

Shepard, R. N. (1992). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 495–532). Oxford, United Kingdom: Oxford University Press.

Tangen, J. M. (2013). Identification personified: Putting people at the core of forensic decision-making. *The Australian Journal of Forensic Sciences, 45,* 315–322. http://dx.doi.org/10.1080/00450618.2013.782339

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science, 22,* 995–997. http://dx.doi.org/10.1177/0956797611414729

Thomas, R., Nardini, M., & Mareschal, D. (2010). Interactions between "light-from-above" and convexity priors in visual development. *Journal of Vision, 10,* 6. http://dx.doi.org/10.1167/10.8.6

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: Experts can do a lot with a little. *PLoS ONE, 9,* e114759. http://dx.doi.org/10.1371/journal.pone.0114759

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in fingerprint identification. *Journal of Forensic Sciences, 58,* 1519–1530. http://dx.doi.org/10.1111/1556-4029.12203

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2014). Human matching performance of genuine crime scene latent fingerprints. *Law and Human Behavior, 38,* 84–93. http://dx.doi.org/10.1037/lhb0000051

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers." *Psychological Bulletin, 76,* 105–110. http://dx.doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. http://dx.doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 108,* 7733–7738. http://dx.doi.org/10.1073/pnas.1018707108

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE, 7,* e32800. http://dx.doi.org/10.1371/journal.pone.0032800

Vokey, J. R., Tangen, J. M., & Cole, S. A. (2009). On the preliminary psychophysics of fingerprint identification. *The Quarterly Journal of Experimental Psychology, 62,* 1023–1040. http://dx.doi.org/10.1080/17470210802372987

Wells, G. L., Wilford, M. M., & Smarlarz, 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. http://dx.doi.org/10.1016/j.jarmac.2013.01.004

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review, 21,* 100–106.

Wulf, G., & Schmidt, R. A. (1989). The learning of generalized motor programs: Reducing the relative frequency of knowledge of results enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 748–757. http://dx.doi.org/10.1037/0278-7393.15.4.748

Young, M., Brooks, L., & Norman, G. (2007). Found in translation: The impact of familiar symptom descriptions on diagnosis in novices. *Medical Education, 41,* 1146–1151. http://dx.doi.org/10.1111/j.1365-2923.2007.02913.x

Young, M. E., Brooks, L. R., & Norman, G. R. (2011). The influence of familiar non-diagnostic information on the diagnostic decisions of novices. *Medical Education, 45,* 407–414. http://dx.doi.org/10.1111/j.1365-2923.2010.03799.x

Appendix A

Experiment 1 Instructions

In this experiment, you are going to take on the role of a fingerprint examiner. You'll start by reading the *Incident/Investigation Report*, which lists the type of crime and a summary of the case. Please read this report carefully. We're going to ask you about the details of the case later in the experiment.

Next, you're going to examine a fingerprint that was lifted from the crime scene and the fingerprint of the suspect. Your job is to determine whether these two prints came from the same person. So carefully analyze the two prints before making your decision. Once you have carefully read about the case and rated the similarity of the fingerprints, you can move on to the next case.

Remember: Read each of the cases carefully because we'll ask you about them at the end of the experiment.

If you have any questions about the experiment, please ask the experimenter now. Otherwise click the "Begin" button below.

Appendix B

Sample Case Reports Presented in Experiment 2: Manipulation of Similarity

Original Case Report Presented During the Learning Phase

"In the present case, the offender entered a shop, which consisted of a small post office and bank agency. At the time, the shop operator and victim in this case was working behind the counter alone and there were no other customers in the shop. After making enquiries with her about the availability of banking facilities, the offender obtained and brought to the counter some bank deposit and withdrawal slips. He asked the victim to fill out a slip. She began to do so, but then the offender said, 'I will stab you with this knife, okay.' The victim looked up, she saw that the man had a knife held to his own throat. The offender demanded money from the shop safe. The victim acceded, opening the safe and removing cash from it while the offender emptied the cash register. The offender ordered the victim to lie down on the floor, and he left the premises with approximately $9,000 cash. A latent fingerprint was found at the scene."

Corresponding Similar Case Report Presented During the Test Phase

"In the present case, the offender entered a shopping center, which consisted of a small pawn store and currency exchange stall. At the time, the store attendant and victim in this case was working in the store by herself and there were no customers in the store. After questioning the victim about cash conversions, the offender brought to the attendant some items for conversion. He asked the victim to exchange these items for cash. As she began to exchange the items the offender said, 'I will shoot you will this gun, got it.' The victim looked up and saw that the man had a gun pointed at his own head. The offender ordered the victim to retrieve money from the cash register. The victim complied, opening the register draw nearest her and removing cash from it while the offender emptied a second cash register. The offender ordered the victim to lie down on the ground behind the counter, and he left the premises with approximately $3,000 cash. Two latent fingerprints were found on the second cash register."

(*Appendices continue*)

Appendix C

Experiment 2 Instructions

Learning Phase

In this experiment, you are going to take on the role of a fingerprint examiner. You'll start by reading the *Incident/Investigation Report*, which lists the type of crime and a summary of the case. Please read this report carefully. We're going to ask you about the details of the case later in the experiment.

Next, you're going to examine a fingerprint that was lifted from the crime scene and the fingerprint of the suspect. Your job is to determine whether these two prints came from the same person. So carefully analyze the two prints before making your decision. Once you have carefully read about the case and rated the similarity of the fingerprints, you can move on to the next case.

Remember: Read each of the cases carefully because we'll ask you about them at the end of the experiment.

If you have any questions about the experiment, please ask the experimenter now. Otherwise click the "Begin" button below.

Test Phase

You have finished the practice section of the experiment. In the next set of trials, you will not be told whether you were correct or not. Again, please read each *Incident/Investigation Report* carefully before examining the prints.

When you are ready to begin, please press the "Begin" button below.

Appendix D

Experiment 3 Instructions

Learning Phase

In this experiment, you are going to take on the role of a fingerprint examiner. You're going to examine a fingerprint that was lifted from the crime scene and the fingerprint of the suspect. Your job is to determine whether these two prints came from the same person. So carefully analyze the two prints before making your decision. Once you have carefully read about the case and rated the similarity of the fingerprints, you can move on to the next case.

If you have any questions about the experiment, please ask the experimenter now. Otherwise click the "Begin" button below.

Test Phase

You have finished the practice section of the experiment. In the next set of trials, you will not be told whether you were correct or not.

When you are ready to begin, please press the "Begin" button below.

Received September 26, 2014
Revision received July 13, 2015
Accepted July 26, 2015 ■