FOCUS ON THE TASK AT HAND: CONTEXTUAL BIAS IN THE FORENSIC EXAMINATION OF HANDWRITING

Marijn Karen Jikke Kouwenhoven

A thesis submitted for the degree of
Doctor of Philosophy
University of Otago
July 2018

ABSTRACT

Forensic science evidence is a crucial part of the criminal justice system, and forensic examiners have long been held in high regard due to their ability to link suspects to crimes in an objective and scientifically rigorous way. Recently, however, concerns have been raised that forensic examiners could be susceptible to contextual bias, in which exposure to information not relevant to the examination could unduly influence their conclusions. Contextual bias has now been demonstrated in a wide range of forensic disciplines (e.g., fingerprints, DNA). The field of handwriting analysis, however, has received relatively little attention. Handwriting analysis is likely to be particularly susceptible to bias, because examiners are asked to opine on a product of a human behaviour that varies widely—even across samples written by the same person. In this thesis, we limited our focus to signature examinations, which involve very small samples of handwriting.

In Study 1, we used a signal detection framework to examine the effects of task-irrelevant contextual information on laypeople's judgements about questioned signatures. In Study 1A, we presented participants with 20 trials comprising one questioned signature and four known signatures. On each trial, participants were asked to indicate whether they thought the questioned signature was genuine or forged. Half of the trials were accompanied by high-biasing contextual information (e.g., stating that the author of the questioned signature had a criminal record); the remaining trials were accompanied by low-biasing information (e.g., stating that the questioned signature was written by a suspect). Contrary to our expectations, the contextual manipulation did not significantly influence participants' ability to discriminate between genuine signatures, nor did it affect their response bias. In Study 1B, we addressed several limitations of Study 1A that could have prevented us from detecting a context effect. This time, we observed that high-biasing contextual information not only increased participants' response bias; it also reduced their ability to discriminate

between genuine and forged signatures. This latter finding suggested that contextual information might influence the process by which participants examined the signatures.

In Study 2, we therefore focused on the signature examination process using a similar paradigm. We explicitly asked participants either to focus on similarities or differences between the questioned and known signatures. We hypothesised that if high-biasing contextual information led participants to use a positive test strategy, then we should be able to eliminate the context effect by forcing participants to identify similarities between the signatures (i.e., to engage in a negative test strategy). This hypothesis was not supported, however, suggesting that efforts to change the signature examination process per se could be insufficient to mitigate contextual bias. As such, it is important to develop practical and effective ways to either limit exposure to contextual information, or to find ways to present contextual information in a way that is least likely to influence examiner's opinions.

For this reason, in Study 3, we turned our attention to practitioners. Specifically, we explored the current state of contextual information management in questioned document examination by conducting interviews with 19 international professional document examiners—both from government laboratories and private practice. As well as canvassing practitioners' views on contextual bias, we sought information about the sources of contextual information that they encounter, the information that is relevant—and not relevant—to their examinations, their methods for reducing the potential for bias, and the perceived barriers to implementing context management strategies. Based on our findings, we provide recommendations for developing and implementing practical and effective context management systems—both for forensic science in general, and handwriting analysis in particular.

ACKNOWLEDGEMENTS

I wish to thank, first and foremost, my supervisor Associate Professor Rachel Zajac for her invaluable advice and guidance during my PhD candidacy. Thank you for your continuous support, encouragement, patience, and for helping me develop my research and writing skills these past three-and-a-half years.

A huge thank you to Dr Niki Osborne. Your enthusiasm and passion for research in this field—as well as life in general—is incredibly inspiring. Thank you for your constant support and feedback, as well as the many brainstorming sessions, e-mails, and skype calls we had whenever I felt at a loss of what to do.

I would also like to thank Professor Bill Thompson for providing me with the opportunity to travel and work overseas. Being able to meet and collaborate with some amazing researches in this field of psychology was an once-in-a-lifetime opportunity and I am incredibly grateful for all the work Bill did to make it happen.

Also, thank you to all the document examiners who agreed to be interviewed. Finding forensic practitioners who are open, and willing, to discuss the issue of bias can be incredibly difficult. When I began working on the NIJ project I had no idea that so many document examiners would volunteer to share their opinions on human error and bias, and I am so grateful to all the document examiners who were willing to talk to me.

Thank you to everyone in the Zajac laboratory. Writing a thesis can feel very lonely at times, but being able to share my experiences with others who were in the same boat made everything a little bit easier. Thank you to Kimberley, Ellen, Andrew, Jacob, Jenn, and Ella for some great chats, advice, and hang outs.

I would also like to thank my friends and family for their continuous support and encouragement throughout my PhD candidacy. Thank you for the dinners, the laughs, and the distractions when I needed them most. Mum, your love and support means the world to me

and I know I would not have been able to finish this thesis without you. Jelle, thank you for being the best big brother I could ask for. And to Sydney; thank you for being there for me despite living halfway around the world.

And finally, to my dad, thank you. You have been—and always shall be—a tremendous source of inspiration for me. You were there for the start of my PhD journey, and it fills me with infinite sadness that you are no longer with us. This one is for you. I hope I made you proud.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER ONE

FOREWORD

On April 26, 1981, a Dallas woman woke up in the middle of the night to find a male stranger lying on top of her. The man threatened her with a butcher's knife, and in the process cut her on the hand, neck, and back. He then sexually assaulted her before fleeing the house. During the investigation, police officers presented the victim with a photographic line-up of potential suspects, and she positively identified a man named Larry Fuller as her attacker. The police collected a rape kit and had it tested for blood group markers that might further identify Fuller as the perpetrator. As a result of this analysis, a forensic analyst concluded that he was unable to exclude Fuller as the source of the biological material. The positive identification by the victim, along with the fact that Fuller could not be excluded, resulted in Fuller being found guilty by a jury. On August 25, 1981, Larry Fuller was convicted of aggravated rape and sentenced to 50 years in prison (Steinback, 2007).

On April 21, 1992, a 22-year-old woman from Muncie, Indiana, was sexually assaulted behind a vacant building. The victim provided a description of her attacker to the police, who then canvassed the area for any individuals who matched the description. One such individual was William Barnhouse. In addition to the victim's identification of Barnhouse, a blood analyst matched genetic markers from sperm left on the victim's jeans with those collected in a rape kit, and concluded that he was unable to eliminate Barnhouse as the source of the evidence. A hair analyst trained by the Federal Bureau of Investigation (FBI) also stated that a single hair found on the victim's body was a definitive match for Barnhouse's hair. Barnhouse was arrested and put on trial for sexual assault. The positive eyewitness identification of Barnhouse, along with the forensic evidence linking him to the crime convinced the jury that Barnhouse was guilty. On December 15, 1992, William Barnhouse was convicted of rape and criminal deviant conduct, and sentenced to 80 years in prison (*State v. Barnhouse*, 1994).

On May 3, 1992, a three-year-old girl was abducted from her home. Two days later her body was found just 450 metres from her home. One suspect emerged: Kennedy Brewer. Police noted that Brewer had been babysitting the girl when she went missing and that there had been no sign of forced entry into the house. Investigators suspected Brewer of raping and murdering the victim in the house before carrying her body to a nearby creek. A key piece of forensic evidence emerged in the case: a multitude of marks on the victim's body. According to a forensic odontologist, the 19 marks were bite marks that had "without a doubt" been inflicted by Brewer. On the March 24, 1995, Kennedy Brewer was convicted of capital murder and sexual battery and sentenced to death (Garrett & Neufeld, 2009).

The three cases described above have two important things in common. The first is that forensic science evidence was instrumental in convicting the defendants. The second is that the forensic examiners' conclusions about the evidence were misleading and incorrect and should not have been admitted in a court of law.

Fortunately for these men, the Innocence Project exists. Founded in 1992 by lawyers Peter Neufeld and Barry Scheck at the US Cardozo School of Law, the aim of this project is twofold. The first aim is to defend convicted individuals who the project has reason to believe are innocent, often through the use of DNA testing. The Innocence Project's second aim is to reform the criminal justice system to prevent future miscarriages of justice. Unfortunately, the organisation's work has made it clear that wrongful convictions like those described above are not a rare occurrence. To date, the Innocence Project has helped to exonerate 358 wrongfully convicted people, who together served a total of 4,926 years in prison (Innocence Project, 2018). Additionally, according to the National Registry of Exonerations (2018) there have been a total of 2,245 exonerations in the United States since 1989, culminating in more than 19,790 years lost due to wrongful imprisonment.

One of the major benefits of organisations like the Innocence Project is that they can identify how people came to be wrongfully convicted in the first place—data that are instrumental in reforming the criminal justice system. To date, one of the biggest contributing factors to wrongful convictions is forensic science error. In fact, this factor is the second most common contributor to identified wrongful convictions (behind eyewitness misidentification at 75%), playing a role in around 45% of exoneration cases (Innocence Project, 2018).

In Larry Fuller's case, the Innocence Project was able use DNA testing to exclude him as the perpetrator. The additional analyses revealed that the forensic analyst involved in Fuller's trial had made a serious error by testifying that he was unable to exclude Fuller as the source of materials. Both Fuller and the victim were non-secretors, meaning that they did not secrete their A/B/O blood type in their bodily fluids. The sample collected from the rape kit contained blood that was Type O—making it impossible for Fuller or the victim to be the source of the blood. Larry Fuller was exonerated on January 11, 2017, having spent 26 years in prison (Innocence Project, 2017b).

Likewise, the Innocence Project helped William Barnhouse by conducting additional DNA testing on the sperm found on the victim's jeans and biological material collected with the rape kit. The results of these additional tests definitively excluded Barnhouse as the source of the evidence. The Innocence Project also argued that the other key piece of evidence—a single hair found on the victim's jeans, described by an FBI-trained forensic analyst as a definitive 'match' to Barnhouse—should not have been admissible in court. Such absolute conclusions about hair samples are not considered by the FBI to be scientifically valid[1] (FBI, 2015). William Barnhouse was released on March 8, 2017, having spent 25 years in prison (Innocence Project, 2017c).

---

[1] In a joint project, the US Department of Justice, the FBI, the Innocence Project, and the National Association of Criminal Defense Lawyers discovered that microscopic hair comparison analysts had provided erroneous testimony or reports in more than 90% of the cases reviewed (FBI, 2015)

The Innocence Project was also instrumental in the exoneration of Kennedy Brewer. A key piece of evidence in Brewer's conviction was a forensic odontologist's testimony that the multitude of marks on the victim's body were made by Brewer's teeth. Further analysis revealed that the marks not only did not come from Brewer, but that they were not bite marks at all (Innocence Project, 2017a). On February 15, 2008, Kennedy Brewer had his convictions overturned after spending 13 years in prison—seven of which were spent on death row.

How do mistakes like these occur? To answer this question, the National Academy of Sciences (NAS) created an independent Forensic Science Committee consisting of representatives of operational crime laboratories, medical examiners, coroners, legal experts, and a diverse group of scientists. The purpose of this committee was to identify the most important issues faced by the forensic community and to provide specific recommendations to address these issues. In 2009, the National Research Council published a major report on the committee's findings.

The overarching conclusion of the NRC (2009) was that very few forensic disciplines have developed reliable measures of the accuracy of conclusions made by forensic scientists. In particular, forensic disciplines that 'identify' a person (e.g. fingerprint or handwriting analysis) had not established reliable statistics on intra-individual and inter-individual variability in their respective fields; in other words, there was insufficient empirical support for the notion that evidence at a crime scene can be used to identify a single person. One of the most pressing concerns raised by the NRC was the potential for erroneous conclusions in forensic science due to bias (see Kassin, Dror, & Kukucka, 2013, for a detailed overview)—especially in forensic disciplines that make subjective judgments on matching characteristics. The report raised the need for a body of research to investigate the source, and effect, of bias so that practical and effective countermeasures can be developed. Notably, the NRC asserted that any such research programmes would greatly benefit from research in cognitive psychology. This thesis represents one response to this call.

CHAPTER TWO

COGNITIVE SHORTCUTS IN DECISION-MAKING

The world around us is extremely complex. Our brains have to perceive, interpret, and act on a constant barrage of stimuli—sometimes in very restricted time periods. When driving a car, for example, we must constantly attend to other cars, traffic lights, pedestrians, and environmental hazards, and make quick decisions accordingly. The way we react to these cues is influenced by the context in which they arise. When a traffic light turns orange as we approach it, for example, we have to make a split-second decision about whether or not to stop. To do this, we have to consider our distance to the traffic light, our speed, and the speed of the cars in front of us. We are also likely to factor other contextual variables into our decision: whether we are in a hurry, what we know about the timing of the lights and the typical driving behaviour at this particular intersection, and the likelihood of getting stopped by police for making a poor decision. How are we able to make such complex decisions in the moment? The cognitive processes involved are considered below.

Schemas

Our brains have developed the capacity to handle large amounts of information by breaking everything down into its simplest forms: schemas. Also known as the "building blocks of cognition" (Rumelhart, 2017, p. 33), schemas are units of knowledge—each relating to one specific object, action, or concept (Rumelhart, 2017; Wadsworth, 2004). Schemas provide a framework for recognising everything we perceive. Schema activation facilitates rapid object recognition (Sun, Simon-Dack, Gordon, & Teder, 2011) and benefits memory processing by guiding our attention to what is most relevant or informative in any given situation (Amato, 1991; Bobrow & Norman, 1975; Fiske & Linville, 1980; Rumelhart & Ortony, 1976; Smith, 1980; Zadny & Gerard, 1974). Schemas can be data-driven through *bottom-up processing*, in which a subschema activates a higher order schema (e.g., the activation of the *tyre* schema leads to the activation of the *car* schema). They can also be theory-driven through *top-down processing*, in which a higher-order schema activates its subschemas based on expectations. According to Rumelhart (2017, p. 41), “the activation of these subschemata derives from a sort of expectation that they will be able to account for some portion of the input data”. For example, if we perceive a three-dimensional, rectangular object that we think might be a car, the *car* schema activates its subschemas of *tyres*, *windows*, *headlights*, to determine the goodness-of-fit of the *car* schema.

Bottom-up processing. People engage in bottom-up processing when they systematically and objectively evaluate incoming sensory information, facts, or evidence (Fazio, 1990; Fazio & Towles-Schwen, 1999). In bottom-up processing, schema activation goes from part to whole—we perceive lower-order units and integrate them into a higher-order schema using data-driven reasoning (Bartels, 2010; Kinchla & Wolfe, 1979; Palmer, 1975; Rumelhart, 2017). For example, on encountering a rose, subschemas of *stalk*, *petals*, and *leaves* activate the higher-order *flower* schema. Further subschemas of *thorns¸ petal shape,* and *rose scent* activate the higher-order schema of *rose*, allowing recognition to occur.

Bottom-up processing might seem like an ideal way of perceiving and interpreting incoming stimuli, as it provide us with an objective perception of our environment. Unfortunately, however, bottom-up processing is often too inefficient to manage the large amounts of information that we encounter. Furthermore, in situations where incoming information is ambiguous or incomplete—trying to read a smudged document, for example—reliance on bottom-up processing is simply not possible. In these situations, people use what is known as *top-down processing* (Bartels, 2010; Kinchla & Wolfe, 1979).

Top-down processing. Top-down processing occurs when we rely on prior knowledge and experience, expectations, and context to interpret and understand incoming information (Bartels, 2010; Fraser-Mackenzie, Bucht, & Dror, 2013; Kinchla & Wolfe, 1979). As well as allowing us to process information more quickly, top-down processing can help us to ‘fill in the gaps’ to deduce information from an ambiguous or incomplete stimulus

(Bartels, 2010; Fraser-Mackenzie et al., 2013; Kinchla & Wolfe, 1979). In Figure 2.1, for example, even though the letters in the word are partially obscured, top-down processes allow us to conclude that the obscured letters are more likely to be *E* and *A* than *F* and *A*.

Figure 2.1. Example of a situation in which top-down processing is beneficial.

Heuristics. While bottom-up and top-down processes act as scripts for perceiving, recognizing, and interpreting information, *heuristics* inform our decision-making processes based on the information available (Anderson, Reynolds, Schallert, & Goetz, 1977; Tversky & Kahneman, 1974). In a restaurant, for example, our restaurant schema provides us with a script for looking at a menu, ordering food, eating it, and paying the bill—but it would not help us decide what food to order. Heuristics are mental shortcuts that reduce cognitive load by giving us a set of rules or strategies to follow based on our prior experience (Risinger & Loop, 2002; Tversky & Kahneman, 1974). Based on the information we access through schemas, we can apply heuristics to inform our decisions about—and interactions with—our environment (Tversky & Kahneman, 1974). Information processing, therefore, becomes low-effort, allowing us to make quick decisions (Bartels, 2010).

Broadly speaking, there are three types of heuristics (Gilovich & Griffin, 2002; Tversky & Kahneman, 1974). *Representativeness heuristics* are used to determine the probability that an object or a person belongs to a certain category. For example, a man who arrives at a busy park for a blind date with a banker might rely on the representativeness heuristic to look for the person who is most representative of the 'banker' category (Tversky & Kahneman, 1974).

Availability heuristics are used to determine the frequency or likelihood of an event occurring, based on how many memories of such occurrences can be retrieved. For example, when asked if there are more words in the English language that begin with the letter *t* or the letter *k*, people often try to answer this question by thinking of as many words that begin with each letter. Because it is easier to think of words beginning with *t*, we would correctly conclude that more words in the English language start with *t*. Put simply, the availability heuristic provides us with the following rule: if many examples of an event can be retrieved from memory, it is likely that such occurrences have a high frequency or likelihood of occurring (Tversky & Kahneman, 1974).

Finally, *anchoring-and-adjustment heuristics* are used to make numerical estimates when uncertain by deciding on a starting point and altering the judgment until a plausible estimate is reached (Epley & Gilovich, 2006; Jacowitz & Kahneman, 1995; Tversky & Kahneman, 1974). As the name suggests, people begin with a reference point and then make adjustments to this reference point until they have reached what they believe to be the correct answer (Tversky & Kahneman, 1974). For example, asking people if sharks kill more or fewer than 10 people per year will give them an initial anchor of 10. If people are then asked "how many people are killed by sharks every year?" their estimate will be biased by this value, probably resulting in a different answer than if the initial question had been "do sharks kill more or fewer than 100 people per year?"

Contextual information. Schemas and heuristics are useful for interpreting our environment and guiding our behaviour—especially when the information available is limited or ambiguous. These processes, however, rely on *context* to operate. Contextual information facilitates cognitive processes by giving meaning to an otherwise ambiguous or complex situation (Anderson et al., 1977; Bransford & Johnson, 1972). In other words, context facilitates comprehension. In a seminal study demonstrating the importance of context on the comprehension and recall of novel material, Bransford and Johnson (1972) read the following ambiguous passage aloud to participants:

> The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavour. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many can easily arise. A mistake can be expensive as well. The manipulation of the appropriate mechanisms should be self-explanatory, and we need not dwell on it here. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. (p. 722)

The passage referred to washing clothes, but the researchers manipulated if and when this information was given to participants. Participants either learnt of the topic of the passage prior to hearing it, after hearing it, or not at all. All participants then indicated how much of the passage they could understand, and were given seven minutes to recall the passage as accurately as possible.

Not surprisingly, participants who were given the topic before hearing the passage had significantly higher comprehension ratings than participants who were told afterwards or not told at all. Furthermore, these participants also exhibited better recall of the passage content. It appears, then, that knowing the topic helped some of the participants create a context that they could use to understand the passage; it also provided a mental framework they could use when recalling it later. Notably, many of the participants in the other two groups reported conducting a mental search for a situation that the passage might relate to.

That is, in the absence of contextual information, they tried to create their own—providing further evidence that we rely on context for meaningful comprehension of incoming information (Bransford & Johnson, 1972; see also Dooling & Lachman, 1971; Pichert & Anderson, 1977).

Schemas and heuristics benefit us by simplifying the complex tasks of categorizing, assessing probabilities, and predicting values into easier and quicker judgment strategies based on past experience (Tversky & Kahneman, 1974). Early work in this area was fundamental to the body of literature on automatic cognitive processes in quick judgments and decision-making. Although initial discussions of heuristics and cognitive shortcuts were met with various critiques and opposition (Cosmider & Tooby, 1996; Gigerenzer, 1991; Lopes, 1991), the theories proposed by Tversky and Kahneman (1974) have nevertheless shaped more recent work on theories of unconscious cognitive processes and decision-making under uncertainty (Gilovich & Griffin, 2002; Kahneman, 2011). Most notably, research on heuristics and top-down processes promoted the idea that people engage in both initial, reflexive evaluations of incoming information in context as well as more considered, rational assessments (Gilovich & Griffin, 2002). This idea lead to the development of an important model of cognitive processing—the *dual-process model*—which suggests that people engage in either controlled or automatic conscious processes (Evans, 2010; Evans & Stanovich, 2013). The research by Tversky and Kahneman (1974) also helped frame discussions and our understanding of how ostensibly rational and logical people could nevertheless make predictable errors in information processing.

Cognitive Bias

Schemas and heuristics allow people to quickly process a large volume of complex information in a short time, but inappropriate use of—and a reliance on—cognitive shortcuts in situations where reasoning requires more in-depth processes can lead to *cognitive bias*.

Cognitive bias is an umbrella term used to describe the phenomenon whereby factors such as past experiences, subjective judgment and interpretation, and outcome expectations or preferences influence the way we seek, interpret, and recall information (Baron, Beattie, & Hershey, 1988; Bartels, 2010; Kunda, 1990; Lundgren & Prislin, 1998).

Heuristics are especially vulnerable to cognitive bias. For example, availability heuristics are used to determine the frequency or likelihood of a phenomenon occurring based on how many memories of such occurrences can be retrieved. Unfortunately, however, the number of memories that are retrieved can bias the subsequent judgment of how likely such a phenomenon is to occur. For example, most people think that the likelihood of being killed by shark is much higher than that of being killed by a cow. Yet in fact cows kill several times more people annually than sharks do (Centers for Disease Control and Prevention, 2018). This error arises because shark attacks tend to be sensationalised in the media and films, whereas cow attacks are not. Consequently, people can retrieve more memories of shark attacks than cow attacks, in turn leading them to believe that the former is more likely than the latter (Kahneman & Tversky, 1996; Tversky & Kahneman, 1974).

Likewise, anchoring-and-adjustment biases have been shown to lead to systematic errors in estimation. For example, Tversky and Kahneman (1974) asked participants to estimate the percentage of African countries who were a member of the United Nations. The participants were given a starting point by spinning a wheel of fortune, asked to indicate whether their starting number was higher or lower than the correct percentage, and then had to give an estimate of the correct percentage. Participants' estimations of the correct percentage were biased by their starting point; those who were given a lower starting number (e.g., 10) gave significantly lower estimates compared to those who were given a higher starting number (e.g., 65). The anchoring effect was not mitigated by rewards for being accurate (Tversky & Kahneman, 1974; see also: Cervone & Peake, 1986, Epley, Keysar, van Boven, & Gilovich, 2004; Northcraft & Neale, 1987; Wright & Anderson, 1989).

As evidenced by the examples discussed above, cognitive bias can manifest due to contextual information. Although contextual information is beneficial for the processing of ambiguous information, it can also lead to biased interpretation of that information. In Figure 2.2, for example, the character in the centre is ambiguous; it can be seen as either the letter B or the number 13. How we interpret the character is highly dependent on the context. Those who process the image vertically are likely to perceive the character as a 13, whereas those who process the image horizontally are likely to perceive the character as a *B* (Fraser-Mackenzie, et al., 2013; Risinger, Saks, Thompson, & Rosenthal, 2002).

Figure 2.2. Example of a situation in which context influences the perception of an ambiguous stimulus (adapted from Fraser-Mackenzie, et al., 2013).

Of course, the consequences of perceiving a 13 versus a B in the figure above are minimal. The same cannot be said, however, for many real-world situations. In fact, in some professional settings, biased decision-making can have disastrous—and life-threatening—consequences; the use of contextual cues can lead people to unwittingly favour one decision over another, leaving them unable to make rational, objective judgments (Lewis, 2016; Nickerson, 1998; Risinger & Loop, 2002).

For example, biased interpretation of information is an alarmingly common occurrence in the medical field (Singh, Meyer, & Thomas, 2014). Although gathering patient information such as chief complaints, appearance, or physical symptoms is usually beneficial in fast-tracking patient care, such information sometimes results in a *diagnostic hypothesis*, in which the search for—and identification of—symptoms can be influenced by the hypothesis held by the medical professional. For example, if a doctor suspects that a patient might have Cushing's disease, and knows that one common symptom is a moon-shaped face, that doctor might interpret a round face as being moon-shaped (Leblanc, Brooks, & Norman, 2002). Such errors can result in delayed treatment—or even the administration of incorrect treatment, with severe consequences for a patient's health. Diagnostic error is the largest contributor to adverse events in patients, and can result in preventable deaths or permanent disability (Baker et al., 2004; Saber Tehrani et al., 2013; Thomas et al., 2000; Zegers et al., 2009). It is estimated that in the US alone, misdiagnoses result in 40,000-80,000 preventable, in-hospital deaths every year (Leape, Berwick, & Bates, 2002; Newman-Toker & Pronovost, 2009).

Yet while medical errors are highly concerning, they are usually identified (and often remedied) soon after the error takes place—usually because the patient's condition either fails to improve or deteriorates (Baker et al., 2004; Leape et al., 2002; Newman-Toker & Pronovost, 2009; Saber Tehrani et al., 2013; Thomas et al., 2000; Zegers et al., 2009). This identification allows decision-makers to learn from their errors and ideally results in systems being put in place to avoid similar errors in the future. In contrast, in some professional settings, there is limited scope for bias-related errors to be identified and remedied. In fact, such an error might even increase the likelihood of a consequence that the decision-maker sees as a mark of success. One such setting is the criminal justice system (Edmond et al., 2017; Fraser-Mackenzie, et al., 2013).

Researchers have identified vast potential for biased decision-making in the criminal context. For example, forming a hypothesis about a crime early on in a case can result in police investigators seeking and interpreting evidence in a way that supports that hypothesis (O'Brien, 2009). Belief in a suspect's guilt makes investigators and jurors more likely to look for—and find—similarities between a facial composite and the suspect's face (Charman, Gregory, & Carlucci, 2009). Police interrogators who are led to believe that a suspect is guilty ask more guilt-presumptive questions, use stronger interrogation techniques, and exert more pressure to obtain a confession (Kassin, Goldstein, & Savitsky, 2003). For these reasons, physical evidence has long been seen as a more definitive form of evidence in police investigations. Such evidence is considered the 'gold standard' against which all other evidence is compared (Kassin, et al., 2013), acting as a failsafe against human error. Yet such a notion misses a crucial point: even when physical evidence is available in a case, the ultimate decision about what that evidence means is still made by a person. In fact, research increasingly suggests that those who work in the area of forensic science are highly susceptible to cognitive bias (Kassin et al., 2013). This issue is discussed in greater detail in Chapter 3.

CHAPTER THREE

BIASED DECISION-MAKING IN FORENSIC SCIENCE

Forensic science is a powerful tool in the criminal justice system. Although the examination of forensic evidence might only be a small step in a large investigative process, the conclusions reached by forensic examiners can be extremely persuasive to fact-finders (DiFonzo & Stern, 2007; Podlas, 2005). In fact, due to popular television shows such as *CSI: Crime Scene Investigation*, many have noted that jurors place a high level of trust in the reliability and validity of forensic science evidence (Kassin et al., 2013; Schweitzer & Saks, 2007; Shelton, Kim, & Barak, 2006). Yet it is becoming increasingly clear that such expectations are out of step with the field. Although it is imperative that the examination and interpretation of forensic evidence is based on objective standards and evidence-based methodologies (Fraser-Mackenzie et al., 2013), there have been growing concerns that aspects of forensic scientists' work leave them particularly susceptible to biased decision-making (National Research Council (NRC), 2009; President's Council of Advisors on Science and Technology (PCAST), 2016).

Indeed, several characteristics of the forensic science process could promote bias. First, by the very nature of their task and the environment in which they work, examiners are frequently exposed to potentially biasing contextual information. Fingerprint examiners, for example, are likely to visit the crime scene and speak with police investigators; they might hear that the suspect has confessed, has a criminal record, or has no alibi. Much of this information is directional; that is, it points the examiner towards a certain outcome (Kassin, et al., 2013), which could in turn lead them to examine and interpret evidence in a way that makes that outcome more likely (Althubaiti, 2016; Bressan & Dal Martello, 2002; Carlson & Russo, 2001; Darley & Gross, 1983; Wason, 1960). For instance, knowledge of a confession could lead the examiner to attribute less significance to differences between a latent (i.e., from the crime scene) and a known (i.e., from the suspect) fingerprint (Kassin, Bogart, & Kerner, 2012; Stevenage & Bennett, 2017).

Second, forensic scientists are required to examine and interpret evidence that often comes with a considerable degree of ambiguity (Budlowe, et al., 2009; Charman, et al., 2009; Kassin, et al., 2013). A fingerprint at a crime scene, for example, could be partial, smudged, or degraded due to time or environmental processes. According to Kunda (1990), top-down processes cannot exert undue influence in the face of clear evidence to the contrary, because reality constrains perception. When evidence is incomplete or degraded, however, the risk of bias increases (Risinger, et al., 2002; Saks, Risinger, Rosenthal, & Thompson, 2003).

Third, most forensic science disciplines are plagued by an absence of objective standards regarding what constitutes a 'match' between a questioned/latent sample and a known sample (Haber & Haber, 2013). Instead, disciplines involving feature-comparisons (e.g., fingerprint examination, questioned document examination, bullet comparisons) utilise an *Analysis-Comparison-Evaluation-Verification* method (ACE-V) (Kerstholt, et al., 2010; Langenburg, Champod, & Wertheim, 2009; Lewis, 2014), which requires examiners to determine whether two pieces of evidence are 'sufficiently similar' to conclude that they share the same source, but gives no set criteria for them to follow. Unsurprisingly, given that the term 'sufficiently similar' has yet to be defined or quantified, researchers have noted a lack of decision consistency both within and between forensic scientists (Haber & Haber, 2013; Kassin, et al., 2013). Instead, forensic examiners compare the features of one piece of evidence with features from another piece, identifying points of similarity or difference until they feel able to conclude whether the two share the same source i.e., a *threshold-based* decision (Fraser-Mackenzie, et al., 2013). Clearly, the more subjective the decision process, the more risk of biased threshold-based decisions. In fact, fingerprint examiners in one study explicitly reported relying on subjective elements such as the "feeling of a match" in making their decisions (e.g., Charlton, Fraser-Mackenzie, & Dror, 2010, p. 389). In other words, the conclusions made by these examiners were based on the fact that their decision threshold had been met through subjective interpretations of the evidence. Unfortunately, such an approach can put examiners at risk of reaching a decision that is in accordance with a desired or expected outcome (Charlton, et al., 2010; Dror & Hampikian, 2011).

The final factor that is likely to predispose forensic examiners to bias is that they rarely receive feedback on the accuracy of their decisions. Feedback is essential for learning as it allows an individual to shape and adapt their behaviour to reduce the frequency of incorrect decisions (Edmond, et al., 2017). The only 'feedback' that forensic examiners receive, however, is the outcome of a case. And, of course, the outcome of the case is likely to be influenced in the direction of the evidence presented by the testifying expert. Therefore, the acceptance of the evidence cannot not serve as a measure of 'correctness' or the examiner's accuracy, because the fact-finder—like the forensic scientist—is not aware of the ground truth (Fraser-Mackenzie, et al., 2013; Park, 2008).

Together, all of the issues discussed above can render the forensic scientist vulnerable to *confirmation bias*: the tendency to interpret ambiguous information in a manner that is consistent with a prior hypothesis, and to ignore—or display more scrutiny towards—information that holds negative implications for an established hypothesis (Ask, et al., 2008; Edmond, et al., 2015; Elaad, 2013; Fraser-Mackenzie, et al., 2013; Haber & Haber, 2013; Kassin, et al., 2013; Kukucka, Kassin, Zapf, & Dror, 2017; Nickerson, 1998; Saks, et al., 2003).

The landmark demonstration of confirmation bias was conducted by Wason (1960). In his 'rule-discovery' experiments, Wason (1960) presented participants with sets of three numbers (*triples*), which conformed to a rule set by the experimenter. Participants had to propose new triples that would help them discover what the rule was; the experimenter told them whether or not their proposed triple conformed to the rule. For example, a participant given the target triple "2, 4, 6" might hypothesize that the rule was "three consecutive even numbers". There are two ways to test such a hypothesis: participants can either conduct a *positive hypothesis test*, in which they propose a triple they believe would fit the rule (e.g., 10, 12, 14); or a *negative hypothesis test*, in which they propose a triple they believe would not fit the rule (e.g., 1, 2, 3). Wason (1960) found that participants were reluctant to engage in negative hypothesis testing, even when it would have revealed more useful information about the rule. This tendency to default to a positive test strategy can produce misleading feedback in support of a hypothesis by neglecting its falsification.

A growing body of research suggests that similar effects can pose a problem in the analysis of forensic evidence. Most notably, exposure to contextual information can lead to biased decision-making in forensic examiners. Such effects have been demonstrated in a wide range of forensic disciplines, including fingerprint examination (Charlton, et al., 2010; Dror, Péron, Hind, & Charlton, 2005; Dror, Charlton, and Péron, 2006; Langenburg, et al., 2009; Osborne & Zajac, 2015; Searston, Tangen, & Eva, 2015), bloodstain pattern analysis (Osborne, Taylor, Healey, & Zajac, 2016a), handwriting examination (Kukucka & Kassin, 2014), forensic anthropology (Nakhaeizadeh, Dror, & Morgan, 2014), and even DNA analysis (Dror & Hampikian, 2011).

In a seminal study in this area, Dror, Charlton, and Péron (2006) explored the influence of contextual information on decision-making in fingerprint analysts. The authors recruited five fingerprint experts and tested them using a within-subjects design. Participants judged a pair of fingerprints that—unbeknownst to them—they had identified as a match five years earlier. To create the expectation that the prints did not match, the experts were told that the pair was from a high-profile FBI case of misidentification (the Madrid bomber case; see Office of the Inspector General [OIG], 2006; Stacey, 2004). Only one of the five experts remained consistent with their original judgment and declared the prints to be a match. Of the other four experts, one changed their decision to 'inconclusive', and the other three directly contradicted their previous decision by declaring a non-match. According to Dror et al. (2006), the study demonstrates that fingerprint experts are susceptible to erroneous decision-making when exposed to misleading contextual information.

While these findings are striking, there are limitations as to what they allow us to conclude. In particular, Dror et al. (2006) based their conclusions on the findings of a one-trial experiment in which exposure to contextual information was not the only variable in play. Previous research has shown, however, that there is considerable intra-expert variability among forensic scientists—that is, even *without* exposure to other information, experts sometimes change their judgments about the same stimuli over time (Dror & Rosenthal, 2008; Ulery, Hicklin, Buscaglia, & Roberts, 2012). Although it seems likely that Dror et al.'s (2006) findings were attributable to misleading information, the absence of a control group means that caution is warranted in the interpretation of their data. Dror and Charlton (2006) addressed this shortcoming in a follow-up study by including control trials in which the prints were not accompanied by contextual information. As expected, some degree of decision change was observed regardless of the contextual manipulation, but the presence of contextual information was associated with more changes, and these changes were disproportionately made in the direction of the information suggesting a match.

Similar findings have emerged in other forensic disciplines. For example, forensic anthropologists—who often are tasked with providing a biological profile of a deceased person by examining their skeletal remains—sometimes rely on visual methodologies that are non-metric and require the examiner to make a variety of subjective judgments (Byers, 2016). To show how these judgments can be shaped by contextual information, Nakhaeizadeh, Dror, and Morgan (2014) presented participants with pictures of real skeletal remains, and asked them to determine the individual's sex, ancestry, and age at death. One group of participants received information suggesting that the deceased was male, one group received information suggesting that the deceased was female, and a control group received no contextual information. Participants' estimation of the individual's sex varied greatly across the three groups. In the control group, 31% of participants determined the remains to be male, and 69% determined the remains to be female. In the 'male' group, 72% of participants concluded that the remains were male, while 100% of participants in the 'female' group concluded that the remains were female. Nakhaeizadeh et al. (2014) proposed that the contextual information resulted in selective attention to specific features of the evidence, and that examiners might have even given more weight to the information than to the physical evidence itself.

Sometimes, the mere context of the criminal investigation process is enough to create context-driven expectations (Kassin, et al., 2013; Lange, Thomas, Dana, & Dawer, 2011). For example, the mere fact that a fingerprint comes from a suspect could result in an assumption of guilt (Ask & Granhag, 2005). Indeed, Lange et al. (2011) argue that characteristics of the legal system can promote questionable interpretations of ambiguous evidence. In their study on the interpretation of auditory evidence, participants heard degraded audio statements from the person they believed to be either the suspect or the interviewer. The statements were all of a benign nature and contained no incriminating words, for example "I got scared when I saw what it'd done to him". Lange et al. (2011) found that the use of the word 'suspect' created context-driven expectations, biasing participants toward hearing falsely incriminating statements in the recording. Participants who had been led to believe the statement came from the suspect were 4.56 times more likely to hear "I got scared when I saw what *I'd* done to him" instead of "I got scared when I saw what *it'd* done to him" [emphasis added]. Also concerning was the finding that participants were as confident in their accurate interpretations as they were in their misinterpretations (Lange, et al., 2011).

Holding expectations of the outcome of an examination, however, is not the only factor that could bias forensic examiners; more subtle forms of context have also been shown to influence decisions. Dror, Péron, Hind, and Charlton (2005) presented participants with 96 pairs of fingerprints and asked them whether or not each pair was a match. In some trials, the prints were preceded by low-emotion or high-emotion crime scene photos. The authors observed that prints accompanied by high-emotion photos were more likely to be declared a match, but only when the prints were ambiguous. When the prints were a clear match or a clear non-match, the additional information had no detectable influence on the participants' decisions. Dror and colleagues proposed that, in the absence of clear bottom-up data, participants exposed to highly emotional context "filled in the gaps" using top-down processes (Dror et al., 2005). Indeed, forensic examiners themselves perceive that emotional context affects their decisions (Hall & Player, 2008).

Follow-up work by Osborne and Zajac (2015) and Zajac, Barrett, Osborne, Hegermann, and Kouwenhoven (2018) controlled for two possible counter-explanations for Dror and colleagues' data: (1) because the photos were presented in order of increasing emotional intensity, it is possible that participants simply made more matches over time; and (2) it is possible that exposure to any contextual information—not just crime-related information—could affect decision-making when evidence is ambiguous. In a series of studies using an adapted version of Dror et al.'s (2005) paradigm, the authors showed that highly emotional photos indeed made participants more likely to declare two ambiguous prints to be a match—but only when those photos were crime-related. Non-crime-related photos that were matched for emotional intensity did not influence participants' decisions.

It is clear, then, that the emotional context alone cannot account for effects like those seen in the research described above. Another possible explanation for the effects of these more subtle, less directional forms of context is that exposure to contextual information can result in *motivated reasoning*. Motivated reasoning influences a person's beliefs and cognitive strategies applied in a given situation depending on what the goal in that situation is, and can be divided into two categories: the motive to arrive at the accurate conclusion (accuracy-driven), or the motive to arrive at a particular, desired, conclusion (direction-driven). If a person is accuracy-driven in a given situation, then they will rely on beliefs and processes that will ensure the correct outcome. If, on the other hand, the person is direction-driven, they will use cognitive strategies that are most likely to result in the desired conclusion (Kunda, 1990). Although both types of motivated reasoning are vulnerable to cognitive bias, accuracy-driven reasoning is less vulnerable than direction-driven reasoning (Kunda, 1990). This is because the motive to arrive at the correct conclusion leads people to engage in more deep and careful cognitive processing. In contrast, the motive to arrive at the desired or expected conclusion makes people susceptible to confirmation bias (Kunda, 1990; Lundgren & Prislin, 1998; Stuart, Windschitl, Smith, & Scherer, 2017).

Although forensic scientists should be accuracy-driven, the nature of their work can render them vulnerable to becoming direction-driven through a need for cognitive closure (NFC)—a term used to describe an individual's desire to reach a definitive conclusion and resolve confusion and ambiguity (Ask & Granhag, 2005; Webster & Kruglanski, 1994). A qualitative study by Charlton et al. (2010) suggests that forensic scientists are motivated to achieve closure. Fingerprint examiners expressed an explicit desire to avoid or resolve ambiguity, and reported feelings of satisfaction or a "buzz" when finding fingerprint features that helped them achieve closure on the case. Of course, in the forensic context, the decision that is most likely to achieve closure is to link the crime scene evidence to the suspect.

It is important to acknowledge that biased decision-making does not always result in erroneous conclusions. In Nakhaeizadeh et al.'s (2014) forensic anthropology study discussed earlier, for example, the skeletal remains had originally been determined as 'probably female'. This means that in the 'female' group, the contextual information likely steered the participants towards what most would consider the 'correct' answer. Regardless of whether the examiners' conclusions were correct or not, however, the data suggest that they used task-irrelevant information to inform their opinion. In other words, the contextual information may have led the participants to the correct conclusion, but for the wrong reasons. It is the job of any forensic scientist to make an independent judgment about the evidence based on the relevant physical evidence alone (Kukucka, 2014). When evidence is presented in court, it is typically assumed by the judge and jury that the evidence is independent, and one piece of evidence is not "tainted" by another (Hasel & Kassin, 2009). Evidence that is not independent can be thought of as double counting, where the 'weight' of an item of evidence is essentially counted twice, resulting in the body of evidence becoming greater than the sum of its parts and making individual items of evidence more convincing to the jury than they should be (Kukucka & Kassin, 2014; Osborne et al., 2016).

In acknowledgement of these issues, the National Research Council (NRC, 2009) has called for the development of research on the effects of cognitive bias in the forensic sciences, especially the development of research programmes on human error and bias in forensic examination. Suggested areas of research include studies that explore the effects of contextual information in forensic examinations, including studies to "determine whether and to what extent the results of forensic analyses are influenced by knowledge regarding the background of the suspect and the investigator's theory of the case" (NRC, 2009, p. 24). The fundamental goal of this thesis is to contribute to this area of research by investigating a forensic discipline that, until recently, has been largely ignored: forensic handwriting examination.

CHAPTER FOUR

FORENSIC HANDWRITING EXAMINATION

The examination and comparison of handwriting[2] for identification is considered to be the oldest of all the forensic sciences (Risinger & Saks, 1996)—even dating back to Roman Law in 539 AD (Koppenhaver, 2007). Handwriting identification is based on the *principle of individuality*, which assumes that no two people write in exactly the same way (Osborn, 1929; Koppenhaver, 2007; Lewis, 2014). The notion of handwriting individuality dates back many centuries and is still accepted in modern society, as demonstrated by the widespread use of signatures as proof of identity or agreement on legal documents (Lewis, 2014).

The primary task of a forensic document examiner[3] is to determine whether two or more handwritten items were written by the same person. Document examiners analyse a wide variety of questioned documents, including land title documents, bank withdrawal slips, medical prescriptions, threat letters, wills, and purchase receipts. The examiner must consider many different aspects of the evidence including, but not limited to, writing pressure, fluency and speed, slant, direction of travel of the writing, consistency of letter formation, and the spatial relationship between letters and words. To do this, examiners independently assess the "pictorial, structural, and line quality features that are perceived to characterize two sets of writing specimens" (Sulner, 2014, p. 1)—a *known* sample and a *questioned* sample—to determine whether the two sets are: (1) sufficiently similar to conclude that they were produced by the same person (i.e., genuine); (2) sufficiently dissimilar to conclude that the two sets of writing were produced by two different persons (i.e., forged); or (3) produced by the same person, but altered (i.e., disguised[4]) (Dyer, Found, & Rogers, 2006).

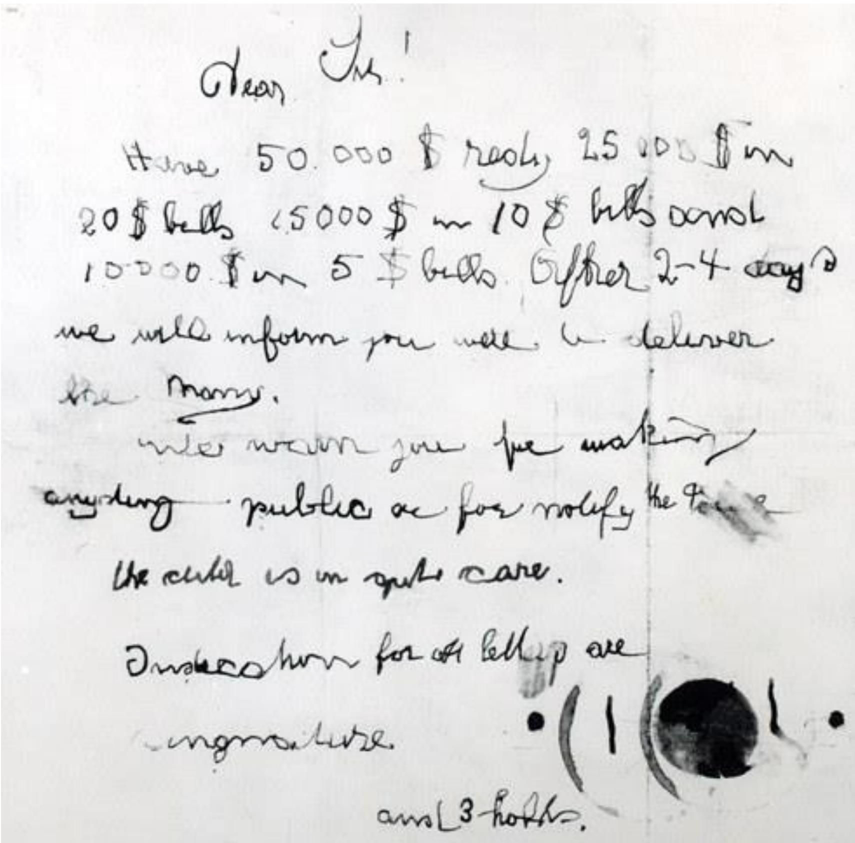
[2] There are many different types of handwriting evidence (e.g., letters, signatures, written names). For the sake of simplicity, I refer to all of these as 'handwriting' unless stated otherwise.
[3] Handwriting examination is one part of the larger discipline of questioned document examination. Forensic document examiners also conduct other examinations, including indented writing examinations, alternate light source examinations, and printing process examinations (SWGDOC, 2018).
[4] Writing that is forged or disguised is sometimes discussed under the umbrella term 'simulated'—a term used to indicate that the questioned writing is unnatural but could still have been produced by the same author (a person might disguise their writing, for example, with a view to denying authorship at a later time).

The analysis of handwriting is particularly challenging because—in contrast to many other forms of forensic evidence—two samples of handwriting from the same author are never identical[5]. That is, all handwriting contains natural variation, such that no one person writes exactly the same way twice (Lewis, 2014; Koppenhaver, 2007; Osborn, 1929). Consequently, whereas the quality of impression evidence (from fingerprints and shoe prints, for example) remains stable, multiple samples of the same handwriting produced by the same author will always vary to some extent (Koppenhaver, 2007).

How, then, do document examiners distinguish intra-writer variation from inter-writer variation? The forensic examination of questioned writing is usually conducted in three phases: *analysis*, *comparison*, and *evaluation* (the ACE method) (Huber & Headrick, 1999; Lewis, 2014; Scientific Working Group for Forensic Document Examination [SWGDOC], 2000;). First, the examiner performs an *analysis* of the questioned writing in the absence of the known writing. Here, the examiner determines the type of writing—assessing its internal consistency, the range of variation, and the presence or absence of any potentially identifying characteristics. Next, following the same steps, the examiner performs an analysis of the known writing, in the absence of the questioned writing. The document examiner then conducts a side-by-side *comparison* of both writings, assessing the comparability of the items of writing, and looking for differences and similarities. Finally, the examiner makes an *evaluation* of the evidence. In this step, the examiner considers the significance of the similarities and differences, and the limitations of the analysis (e.g., missing characters)—both individually and in relation to the rest of the writing. For example, a document examiner has determined that the letter *E* looks different between the questioned and known writing, but also needs to consider the significance of this difference in relation to the rest of the

[5] In fact, if two pieces of handwriting look identical, this is typically indication of forgery using a tracing method (Lewis, 2014).

writing. If only the *E* looks different, then it is not significant enough to conclude that the questioned and known writing were written by different persons. If, on the other hand, the *E* looks different, but so do several other letters, this accumulation of differences could lead to an overall opinion that the writer of the questioned document did not write the known document (Lewis, 2014; Osborn, 1929; SWGDOC, 2000). The examiner then forms a conclusion ranging from *identification* (i.e., "John Doe wrote the questioned material") to *inconclusive* (i.e., "I am unable to determine whether John Doe wrote the questioned material") to *elimination* (i.e., "John Doe did not write the questioned material"). (SWGDOC, 2000). See Figures 4.1 and 4.2 for an example of handwriting examination in practice.



Dear Sir!
Have 50.000 $ redy 25 000 $ in
20 $ bills 15000 $ in 10 $ bills and
10000 $ in 5 $ bills After 2-4 days
we will inform you were to deliver
the mony.
We warn you for making
anyding public or for notify the police
the child is in gut care.
Indication for all letters are
singnature
and 3 holds.

Figure 4.1. A copy of a ransom note, taken from the FBI files (FBI, n.d.).

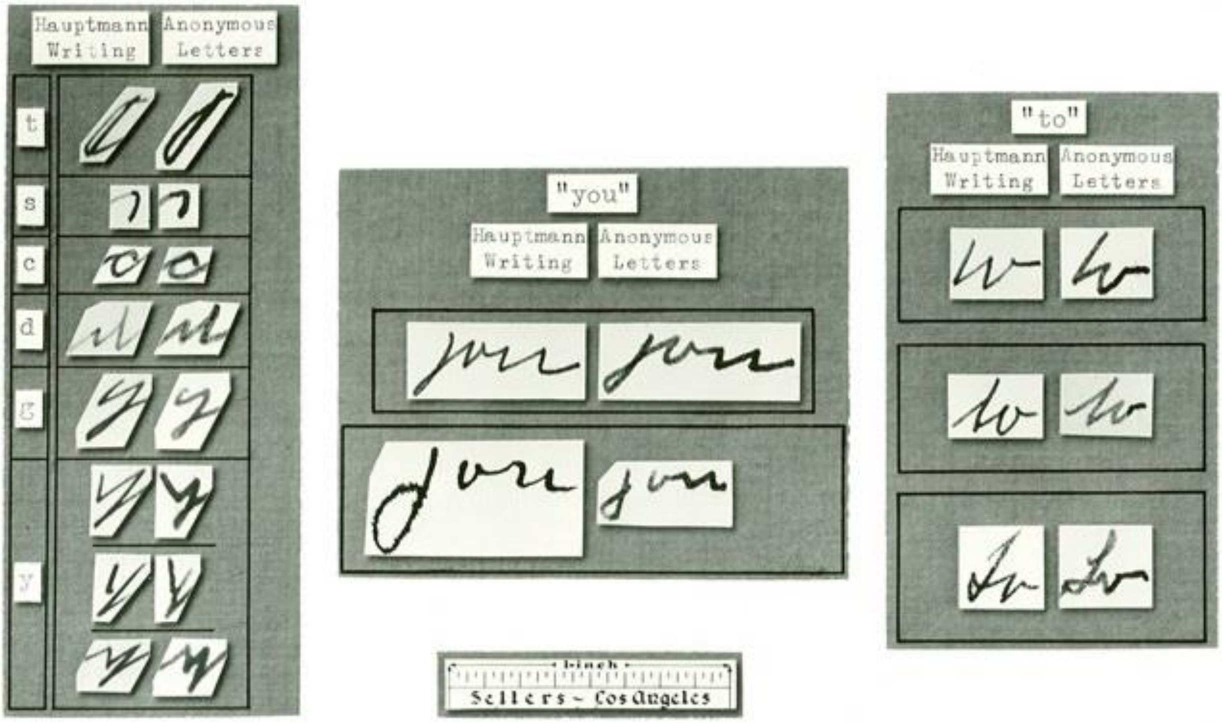


Figure 4.2. Charts used by FBI handwriting experts to compare a suspect's writing to writing on the ransom note (FBI, n.d.).

Validity and Reliability of Forensic Handwriting Examination

Although the forensic examination of handwriting has existed as a discipline for centuries, it is a scientific field that nonetheless lacks a strong empirical foundation (NRC, 2009; Risinger & Saks, 1996). Traditionally, those who worked in this field were self-proclaimed experts, whose expertise was attributed to their experience outside of the criminal justice process—for example, postal inspectors and bank tellers, who scrutinised handwriting daily (Risinger & Saks, 1996).

In contrast, today's document examiners learn the skills required for their line of work through mentoring, textbooks and journal articles, professional development workshops, and practical experience (Bird, Found, Ballantyne, & Rogers, 2010a). The American Board of Questioned Document Examiners (ABQDE) maintains standards of qualification, certifying applicants who comply with ABQDE requirements. Certification by the ABQDE applies to residents of the US, Canada, Australia, and New Zealand. Applicants must have a university degree and have completed a full-time training period of at least two years in a forensic laboratory recognized by the ABQDE. The applicant's training programme must also meet the basic requirements described by SWGDOC, they must be actively engaged in the practice of forensic document examination, and they must demonstrate a record of appropriate professional activity (ABQDE, 2014).

Despite the current stringent criteria for becoming a document examiner, some academics still question the validity of forensic handwriting examination and its purported experts. Risinger and Saks (1996), for example, asserted that forensic handwriting examination is not a science due to its static nature—its principles have not changed in nearly a century, with modern document examiners still relying on techniques described by Osborn in 1929. These authors also argued that the discipline suffers from a lack of organized reporting and publications of observations for empirical testing—that is, there is no "agreed taxonomy of sufficient refinement to yield dependably quantified data, or dependably comparable observations of any refinement" (Risinger & Saks, 1996, p. 38).

The concerns raised by Risinger and Saks (1996) led several researchers to conduct studies testing the professional forensic document examiners' skills in a controlled setting (see Table 4.1 for a summary of results). Kam, Fielding, and Conn (1997), for example, asked document examiners and lay people to compare pairs of documents and to determine whether they had been produced by the same person. Lay people were six times more likely to incorrectly conclude that two pairs of documents had been written by the same person. And although the authors did not provide data on this point, they did note that document examiners were much more likely than lay people to conclude that there was insufficient evidence to determine authorship, thereby preventing them from erroneously matching items from two different authors.

Table 4.1.
Results of Studies Investigating Handwriting Comparison Abilities of Document Examiners and Laypeople.

Study	*N* participants	*N* trials	Type of questioned writing	% Correct		% Incorrect		% Inconclusive	
				QDEs	Laypeople	QDEs	Laypeople	QDEs	Laypeople
Kam et al. (1997)	105 QDEs 41 laypeople	144	Forged handwriting Genuine handwriting	-	-	6.5%	38.3%	-	-
Kam et al. (2001)	69 QDEs 10 laypeople	6	Genuine signature Forged signature	85.9% 96.1%	70.0% 92.0%	7.1% 0.5%	26.1% 6.5%	7.1% 3.5%	4.3% 1.4%
Sita et al. (2002)	17 QDEs 13 laypeople	150	Signature	54.8%	57.1%	3.4%	19.3%	41.8%	23.6%
Dyer et al. (2006)	9 QDEs 12 laypeople	32	Signature	77.8%	45.3%	-	-	-	-
Found & Rogers (2008)	Unavailable	29,811	Genuine signature Forged signature Disguised signature	85.0% 45.3% 26.9%	-	2.2% 3.4% 18.0%	-	12.8% 51.3% 55.1%	-
Bird et al. (2010b)	11 QDEs 10 laypeople	140	Handwriting	73.8%	80.1%	3.4%	11.4%	23.5%	8.4%
Kam et al. (2015)	19 QDEs 26 laypeople	8	Genuine Handwriting Forged handwriting	94.3% 81.7%	66.8% 57.2%	0.0% 10.0%	17.6% 32.7%	5.7% 8.3%	15.4% 10.0%

Note. QDE indicates Questioned Document Examiner. "-" indicates the authors did not provide the relevant data.

Bird and colleagues (Bird, Found, & Rogers, 2010) compared how well document examiners and lay people could distinguish between natural and disguised handwriting. The two groups displayed very different response profiles, with document examiners making significantly more inconclusive judgments. When they did give an opinion, document examiners were significantly more accurate than lay people (Bird, et al., 2010b). This pattern of findings is consistent with those of other studies, including those examining signature comparisons (Kam, Gummadidala, Fielding, & Conn, 2001; Sita, Found, & Rogers, 2002).

According to Bird et al. (2010a) and Kam et al. (2001), the expertise of document examiners lies in their significantly lower error rate relative to lay people. Because lay people are less likely to make inconclusive decisions, however, they are also more likely to accrue more incorrect opinions. It can be argued, then, that document examiners outperform lay people because the former are better at recognising that they do not have sufficient information to complete the task (Towler et al., 2018).

To examine this possibility more closely, Sita et al. (2002) analysed their data in terms of 'called' opinions—that is, trials in which a participant did not make an inconclusive decision. According to Sita et al. (2002, p. 5), "called opinion rates represent correct or error rates when subjects were prepared to express an opinion other than inconclusive and are *arguably the rates with most significance for legal determinations*" [emphasis added]. In called opinion trials, the document examiner group exhibited an error rate of 5.8%, compared to 25.3% in the lay sample (see Kam, Abichandani, & Hewett, 2015, for similar findings). That is, even when inconclusive decisions were omitted, document examiners made fewer errors than lay people. Notably, Sita et al. (2002) also found that document examiners attained lower error rates when the signatures were simulated, relative to when they were genuine. This pattern of findings is likely due to document examiners preferentially interpreting points of difference as evidence of simulation, rather than to the natural variation in the writing. Interestingly, document examiners made more correct decisions when signatures were high—as opposed to medium—in complexity. According to Sita et al. (2002), there is logic to the notion that it is easier to spot a forgery when a signature is complex; the more complex a signature is, the more difficult it is for a simulator to reproduce it.

There appears, then, to be some validity to the purported expertise of document examiners, but it remains unclear what exactly makes document examiners experts. Previous research in other domains suggests that expertise offers an advantage in the use of visual information to search for important features that aid in problem-solving (Knoblich, Ohlsoon, & Raney, 2001; Reingold, Charness, Pomplum, & Stampe, 2001; Russo, Pitzalis, & Spinelli, 2003). According to Dyer et al. (2006), the human examination system relies on two components: the cognitive system, which makes decisions about the significance of features in the evidence; and the visual system, which is used to search for important features in the evidence. The studies discussed previously focused on the cognitive system and the final opinion given by participants, essentially treating participants like a black-box, with visual images as the input and opinions as the output. Dyer et al. (2006), however, argue that to fully understand document examiner expertise, it is necessary to investigate what happens in the black-box *during* the examination. Therefore, an alternative method for investigating how experts differ from lay people involves examining visual attention.

Dyer et al. (2006) designed a visual attention experiment to investigate whether document examiners and lay people differ in the process by which they examine questioned signatures. More specifically, the authors sought to discover what features of a signature document examiners and lay people pay attention to when determining authorship. Using eye-tracking software to measure visual attention during a questioned signature comparison task, the authors found that there were no behavioural differences in the way that document examiners and lay people devoted visual attention to signature features. Dyer et al. (2006) therefore suggested that, rather than a superior visual search for features in the signature, the superior accuracy of the document examiners reflects an enhanced cognitive system and improved processing of the visual features in the signatures—likely developed through training and experience in the field.

Although the evidence suggests that document examiners are better than lay people at forensic handwriting examination, it is important to note that document examiners are by no means immune to error. Found and Rogers (2008) used a novel method to investigate potential contributors to document examiners' error rates: they inserted experimental trials into document examiners' daily workflow, resulting in a total of 29,811 decisions made over a five-year period. Close inspection of Found and Rogers' (2008) data revealed that the rate of inconclusive and incorrect decisions differed as a function of the nature of the questioned signatures. Specifically, document examiners were less likely to make mistakes on questioned signatures that were genuine or forged, relative to questioned signatures that were disguised. More than half of the opinions on disguised signatures were inconclusive, but when document examiners *did* give an opinion it was incorrect 40% of the time. According to Found and Rogers (2008), when a signature is disguised or complex, document examiners have difficulty deciding which alternative authorship explanation is the most probable cause for the observed combinations of similar and dissimilar features. When experiencing such difficulties, document examiners will default to attributing any observed dissimilar features to differing authorship rather than to intra-writer variation.

Although such a strategy might result in a higher accuracy rate when a signature is forged, it could increase errors when a signature is genuine, especially if a document examiner has been exposed to case-related contextual information (Kukucka & Kassin, 2014). As discussed in Chapter 3, exposure to contextual information could increase forensic scientists’ motivation to solve a case (Charlton et al., 2010) or create expectations about a suspect’s guilt (e.g., Dror & Charlton, 2006; Nakhaeizadeh et al., 2014). In combination with a tendency to mistake intra-writer variation for differing authorship, exposure to contextual information could place document examiners at particular risk of making the very error that is most likely to result in a miscarriage of justice. The overarching goal of this thesis is to investigate contextual bias in handwriting examinations, and specifically in the examination of signatures.

CHAPTER FIVE

OUTLINE OF THE RESEARCH

Forensic scientists operate under the assumption that, through a thorough examination of the evidence, they can identify the source of that evidence. Recent research, however, has called into question the validity and reliability of many of the forensic disciplines, including fingerprint examination (Charlton et al., 2010; Dror et al., 2005; Dror et al., 2006; Langenburg et al., 2009; Osborne & Zajac, 2015; Searston et al., 2015), bloodstain pattern analysis (Osborne et al., 2016a), handwriting examination (Kukucka & Kassin, 2014), and even DNA analysis (Dror & Hampikian, 2011)—which is generally considered the 'gold standard' against which all other disciplines are measured (Lieberman, Carrell, Miethe, & Krauss, 2008). One reason for this shift in attitude towards forensic science is the growing concern about forensic scientists' vulnerability to cognitive biases, whereby factors inherent in the investigation process result in undue reliance on top-down processing and, in turn, biased interpretations of the evidence. The field of handwriting analysis is likely to be particularly susceptible to bias, because examiners in this field are asked to opine on a product of a human behaviour that varies widely—even across samples written by the same person (Koppenhaver, 2007; Kukucka & Kassin, 2014).

In this thesis, we[6] explored the effects of contextual information on the examination of handwriting, with a focus on signature comparisons. Signatures are unique and highly specialised items of handwriting (Park, 2008). There exists only a small amount of comparable material for any questioned signatures which makes it challenging to attribute significance to similarities and differences (Giles, 2004). In fact, signature analyses and comparisons is one of the most challenging and difficult areas of document examination (Bird et al., 2010; Osborn, 1929; Park, 2008). Bias is more likely to manifest in cases where the data are limited or ambiguous (Kassin et al., 2013; Kunda, 1990; Lange et al., 2011),

[6] While the work in this thesis is my own, I received advice and feedback from my supervisor and a postdoctoral fellow. I use the term "we" to reflect these facts.

especially if contextual information supporting one outcome over the other is present (Ask et al., 2008; Dror et al., 2005; Page et al., 2012). As such, document examiners may be especially vulnerable to contextual bias in signature comparisons.

The thesis comprises two laboratory-based experiments investigating context effects in laypeople's examinations of questioned signatures, and one field study using qualitative methods to explore questioned document examiners' views on contextual bias and its management. The purpose of the laboratory-based experiments was to develop signal detection paradigms that could eventually be used to test for context effects in forensic examiners. Using signal detection frameworks to explore contextual bias in forensic examiners is a novel approach, and as such we thought it best to test these paradigms on laypeople first.

Chapter 6: Can Contextual Information Influence Laypeople's Determinations of Genuine and Forged Signatures?

Handwriting examiners rarely examine questioned handwriting in the absence of potentially biasing contextual information. For example, they often have access to the details of the offence, and might also be informed about the criminal history of the suspect, whether that suspect has confessed, and other laboratory results. Furthermore, the environment in which the handwriting examiner works (e.g., a police laboratory) has inbuilt expectations that the suspect is guilty—increasing the risk that the handwriting examiner will opine that the questioned sample is forged.

In **Study 1A**, we used a signal detection framework to examine how these kinds of context might affect the *outcome* of a signature examination. To do this, we presented participants with a series of questioned signatures, each accompanied by four known signatures. In each trial, participants were required to determine whether the questioned signature was genuine or forged. In a within-subjects fashion, we manipulated whether trials were accompanied by contextual information suggesting that the signature was forged.

In **Study 1B**, we changed the dichotomous ‘forged’ or ‘genuine’ response option into a more sensitive forced-choice confidence rating of ‘sure genuine’ to ‘guess genuine’ and ‘guess forged’ to ‘sure forged’. This allowed us to measure our participants’ confidence in their decision. We also changed the contextual information so that it could be presented directly alongside the questioned and known signatures.

Chapter 7: Can the Way that Laypeople Evaluate Questioned Signatures Change the Effect of Contextual Information?

In Chapter 7, we considered how contextual information might affect not only the examination outcome, but also the examination *process*. Although Sulner (2014) suggested several ways in which handwriting examination processes might become biased, studies have yet to investigate the effects of contextual information on the handwriting examination process (Kukucka., 2014). The aim of **Study 2** was to fill this gap in the literature.

As in Study 1, we presented participants with a series of questioned and known signatures, some of which were accompanied by information that contained directional cues to suggest a forgery. Prior to determining whether the questioned signatures were forged or genuine, however, participants were required to identify points of similarity or difference between the questioned and known signatures. In this way, we could assess the influence of contextual information on the way in which lay people make their decisions.

Chapter 8: What are Questioned Document Examiners’ Perspectives on Managing Contextual Bias?

Academic commentators have suggested several strategies for mitigating cognitive bias in forensic science. Although academics’ efforts on this issue have been commendable, those efforts have typically taken place in the absence of meaningful consultation with stakeholders. Current recommendations in the literature, for example, do not take into account the practical constraints of forensic work; instead, they focus on how context management procedures might be effective in ideal scenarios. Recommendations in this area have a 'one size fits all' approach, neglecting the fact that not all context management strategies are feasible for all forensic disciplines, or for both institutional and privately-run organizational structures. This disparity between academics' and practitioners' perspectives has led to a reluctance from the forensic science community to adopt—or even consider—such recommendations (Budlowe et al., 2009; Budlowe, 2010; Butt, 2013; Evett, Berger, Buckleton, Champod, & Jackson, 2017; Thornton, 2010). In particular, there is uncertainty and disagreement about: (1) what contextual information might be relevant/irrelevant to the forensic scientist; (2) the circumstances under which context management is necessary; and (3) the specific context management procedures that will be realistic and effective in practice.

In **Study 3**, we took a closer look at these issues by examining the current state of contextual information management in forensic handwriting examination. The broad aim of this study was to discuss the issue of cognitive bias with questioned handwriting examiners, to determine what types of contextual information examiners consider to be task-relevant, and to seek input regarding which context management strategies are—and are not—fit-for-purpose.

CHAPTER SIX

STUDY 1: CAN CONTEXTUAL INFORMATION INFLUENCE LAYPEOPLE'S DETERMINATIONS OF GENUINE AND FORGED SIGNATURES?

Achieving objectivity is particularly difficult for the handwriting examiner because of handwriting's "dynamic function" (Miller, 1984, p. 407). While evidence such as DNA and fingerprints have stable features that can be defined prior to analysis (e.g., the rarity of certain minutiae, or the likelihood of certain alleles corresponding; Cappeli, Ferrara, & Maltoni, 2012; Luftig & Richey, 2001), the features of importance or uniqueness in a handwriting sample cannot be predefined (Osborn, 1929). Furthermore, while fingerprints and DNA remain stable over time, a person's handwriting varies due to myriad factors, including aging, the writing surface, medication, or writing pace. Even the same sentence copied several times in the same time period and under the same conditions can show considerable variation (Morris, 2000). For the handwriting examiner, distinguishing between inter- or intra-writer variability is therefore crucial (Koppenhaver, 2007; Lewis, 2014). Another troublesome factor is a lack of institutional information providing an objective reference point for the rarity or commonality of individual handwriting features, meaning that the main point of reference will always be the examiner's own experience (Risinger & Saks, 1996). The resulting subjectivity in the comparison process (Miller, 1984; Sulner, 2014) leaves examiners vulnerable to task-irrelevant factors that could bias their decisions (Sulner, 2014).

Cognitive bias in handwriting analysis can manifest in several ways, most of which are due to exposure to contextual information. According to Found and Ganas (2013), there are three primary sources of bias-inducing contextual information in this field. The first is the context of the examination. As with many forensic science disciplines, handwriting analysis is often conducted within policing institutions, where investigators typically submit known materials of only one person—the suspect—for examination. Found and Ganas (2013) suggested that such a system leads to an inbuilt expectation of the suspect being the writer of the questioned document, because the known sample always comes from an individual the police believe is guilty (Lange et al., 2011).

The second potential source of contextual information is the documentation related to each case. Specifically, paperwork that accompanies evidence submitted for examination often contains information that is task-irrelevant for the document examiner. This might include, for example, non-directional cues that create a motivation to solve the case (e.g., the nature of the alleged offence) or an explicit expectation of guilt (e.g., the suspects' criminal history; Found and Ganas, 2013).

The final potential source of contextual information in handwriting analysis is face-to-face communication. Because most forensic science analyses are conducted within policing institutions, conversations between investigators and forensic examiners are common, and can lead to the transmission of task-irrelevant information (e.g., that the suspect has confessed to the crime; Saks et al., 2003). Similar exchanges can also happen via telephone or email (Found & Ganas, 2013). Examiners might also be made aware of what their colleagues have previously concluded when asked to review their findings (Kassin et al., 2013).

How might these factors influence the examination of handwriting? Miller (1984) conducted the seminal study aimed at answering this question, by asking two groups of document examiner trainees to examine and compare handwriting samples for a criminal investigation. Both groups were given a summary of facts about the investigation, the suspect's name and exemplars of his handwriting, and three items of questioned writing found on suspicious cheques. Miller (1984) informed participants in Group 1 that the handwriting exemplars came from the suspect, and that two witnesses reported having seen the suspect write out and pass the cheques. Participants in Group 2 were not given this additional information; instead, they were given two additional comparison items from two additional suspects and asked to examine the comparisons provided by all three suspects to determine if any of them wrote any of the questioned cheques. In reality, none of the questioned cheques had been written by any of the suspects. Four of the six examiners in

Group 1, however, concluded that the handwriting on the questioned cheques matched the exemplars of the suspect's handwriting; one participant reported being unable to make a decision and only one participant correctly reported a non-match. In contrast, all six examiners in Group 2 correctly reported that none of the suspects wrote the questioned cheques (Miller, 1984). According to Miller (1984), these findings provide the first piece of empirical evidence to suggest that the analysis and comparison of handwriting is susceptible to context effects.

More specifically, Miller (1984) posited that his findings were highly specific to the nature of handwriting analysis. He noted that, in any handwriting comparison, it is always possible to note both similarities and dissimilarities in the evidence. If the samples are very similar, then it becomes easier to note similarities, while dissimilarities are less obvious. Miller (1984) proposed that handwriting samples that are highly similar might serve to enhance any pre-existing bias towards a match (i.e., the suspect wrote the questioned writing), and lower the threshold for making that conclusion. He noted that it was even possible that the similarity between the handwriting samples provide further enhancement of the participants' expectation that the suspect wrote the questioned document.

However, another possible explanation for Miller's (1984) findings is that the two experimental groups used different methods of comparison. Unlike Group 1, who saw only one exemplar from the suspect, Group 2 examined handwriting samples from three different suspects. The latter group essentially carried out an *evidence line-up*—one of the methods that has been proposed to increase objectivity in forensic science (Saks et al., 2003; Wells, Wilford, & Smalarz, 2013) (See Chapter 8, for a detailed discussion). It is therefore conceivable that the method of comparison—rather than the absence of contextual information— is accountable for Group 2's better performance.

Kukucka and Kassin (2014) addressed this by investigating the potential for confessions to influence perceptions of handwriting evidence—even when the comparison method was held constant. When lay participants were told that the defendant had confessed to a bank robbery—even though the confession was later recanted due to claims of coercion—they were more likely to incorrectly implicate the defendant as the author of a hold-up note. In a follow-up study, the authors asked laypersons to judge the same handwritten notes at two different time points. Participants were shown eight pairs of handwriting samples; they were asked to rate each pair on similarity and indicate their confidence that the samples matched. Between five and nine days later, participants were assigned to one of three conditions: *confession-present*, in which they read a case summary that mentioned a recanted confession; *confession-absent*, in which they read a case summary that did not mention a recanted confession; or *control*, in which they did not receive any case information. Participants were then asked to judge one pair of handwriting samples they had examined earlier. Participants in the confession-present condition were much more likely to conclude that the handwriting samples matched—that is, that the defendant wrote the hold-up note—at Time 2 than at Time 1. In contrast, participants in the confession-absent and control conditions showed no significant change in judgments over time (see Figure 6.1). Participants in the confession-present condition were also more likely to judge the defendant as guilty (43%) than participants in the confession-absent condition (5%).

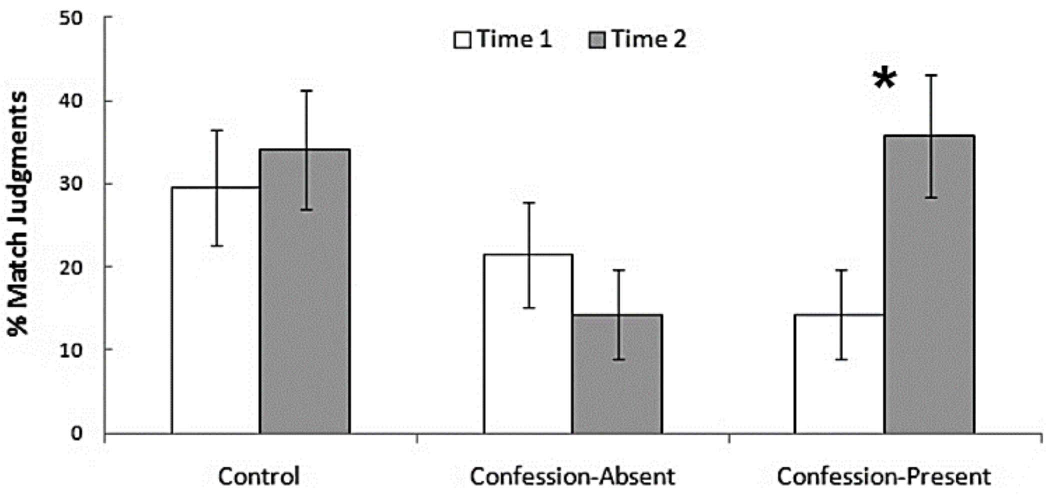


Figure 6.1. Kukucka and Kassin's (2014) data showing mean percentage of match judgments as a function of time and confession. (Adapted from "Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias", by J. Kukucka and S. Kassin, 2014, *Law and Human Behavior, 38* (3), p. 265. Copyright 2013 by American Psychological Association.)

Unpublished data also suggest that handwriting examinations could be susceptible to context effects. In a report to the U.S. Department of Justice, Merlino (2015) described a study investigating the effect of contextual information on signature examinations. In particular, the study focused on the potential biasing effect of knowing the initial examiner's conclusion—knowledge that could steer the reviewer's examination towards the same outcome (Saks et al., 2003). The experimenters took signatures that individual examiners and laypeople had determined to be genuine, disguised, or simulated; they were later asked to reconsider the signatures under the guise of reviewing another examiner's conclusions. That is, they were oblivious to the fact that they were actually re-examining their own work. The study found that participants showed a tendency for participants to change their original conclusion to conform to that of the previous examiner.

Taken together, the findings discussed above suggest that contextual information can exert a considerable influence on handwriting comparisons. Previous research on context effects in forensic science, however, has focused simply on how context affects the number of match and non-match decisions made. Although there is evidence to suggest that context is associated with an increase in match decisions (Dror et al., 2005; Osborne & Zajac, 2015), what remains unclear is *how* contextual information affects forensic decision-making (Phillips, Saks, & Peterson, 2001). A signal detection approach to data analysis could provide new insights into the underlying mechanisms involved (Phillips et al., 2001; Searston et al., 2015; Thompson, Tangen, & McCarthy, 2013).

Signal Detection Theory (SDT). The signal detection framework is often used to examine how individuals make binary decisions (Phillips et al., 2001). This framework is applicable in any situation in which two stimuli need to be discriminated—for example, a genuine signature from a forged one—and is especially useful in situations involving a considerable element of uncertainty (Stanislaw & Todorov, 1999). In a typical SDT paradigm, participants are asked to determine whether a signal is present among noise. In a lie detection task, for example, the polygraph administrator must decide whether the test taker is lying (signal present) or telling the truth (signal absent), but the signal can be obscured by artefacts (e.g., the test-taker moving; Stanislaw & Todorov, 1999).

There are four possible outcomes in an SDT paradigm. A *hit* occurs if the observer correctly states that the signal is present; a *false alarm* occurs if the observer states that the signal is present, but it is actually absent. A *correct rejection* occurs if the observer correctly states that the signal is absent; a *miss* occurs is if the observer states that the signal is absent, but it is actually present.

These four response categories can be used to obtain two outcome measures: sensitivity and response criterion. Sensitivity (*d*') is a measure of the ability to determine whether a signal is present or absent, and depends on both the observer's diagnostic ability and the quality of the stimulus (Stanislaw & Todorov, 1999). The sensitivity calculation compares performance on noise-only trials and signal-and-noise trials, and is calculated using the following formula, (where FA refers to false alarms, and CR refers to correct rejections):

$$d' = z(Hits/(Hits+Misses)) - z(FA/(FA+CR))$$

A *d*' of 0 means that the individual is unable to distinguish signals from noise. Note that *d'* measures the distance between the signal and noise-only means in standard deviation units, making it difficult to interpret specific values other than zero. Furthermore, there is no upper limit for the value. That said, the higher the value of *d'*, the better the individual is able to distinguish the signal from noise (Stanislaw & Todorov, 1999).

Response criterion (*c*) is the value given to an individual's decision threshold. In an SDT paradigm, the participant decides whether the signal is present or absent based on a subjective value that the observer assigns to the stimulus on each trial. If the decision-maker determines that the threshold is met (i.e., the response criterion is achieved), they will make a signal-present response; if it is not met the decision-maker will make a signal-absent response (Stanislaw & Todorov, 1999). Response criterion (*c*) is calculated using the following formula:

$$c = (d'/2) - z(Hits/(Hits+Misses))$$

Response criterion can shift depending on a person's response bias. The degree of response bias affects how likely a person is to state whether the signal is present or absent when the person is unsure—for example, when the noise-to-signal ratio is large. In signal detection theory, bias refers to an overall tendency to favour one response over the other (Klayman, 1995). The response bias measure can range from -1 (conservative response bias; the observer favours "signal absent" responses) to +1 (liberal response bias; the observer favours "signal present" responses). A *c*-value of 0 indicates no response bias—that is, when unsure, the individual is equally likely to conclude that the signal is present or absent (Stanislaw & Todorov, 1999).

Accuracy is a function of both the examiner's diagnostic ability to discern the signal from noise (sensitivity) and the decision-threshold (response criterion) used to determine whether or not a signal is present. While an examiner's sensitivity typically remains relatively constant (Phillips et al., 2001), the response criterion can change for each stimulus due to factors such as knowledge of contextual information. For example, consider a fingerprint examiner who has been asked to compare a smudged fingerprint recovered from a crime scene to a suspect's fingerprint, and make a judgement as to whether they match (signal present) or do not match (signal absent). If the examiner is informed that the suspect confessed, this knowledge could lower the examiner's decision threshold. That is, although the examiner's sensitivity has not changed, they would be more likely to conclude that the fingerprints match due to a shift in response criterion (Phillips et al., 2001).

SDT is a useful method of testing performance in forensic examiners because it allows researchers to measure sensitivity independent from response criterion (Phillips et al., 2001). In doing so, it is possible to determine whether contextual bias is due to a change in sensitivity, a shift in response bias, or a combination of both (Searston et al., 2015; Thompson et al., 2013). In the first study to use SDT to investigate context effects in a forensic decision-making task, Searston and colleagues (2015) presented laypeople with pairs of fingerprints that were either from the same or different sources. Each pair was accompanied by case reports and related images of either a "severe" (e.g., homicide) or "not-severe" crime (e.g., theft) (p. 52). Participants were more likely to declare two fingerprints to be a match when the case information was severe compared to not-severe, but overall sensitivity did not change as a function of contextual information. In other words, participants were capable of discriminating matching and non-matching fingerprints equally well on severe and non-severe trials. Response criterion, however, was affected. Participants showed a more liberal response bias (a higher tendency to conclude that the prints matched) when presented with severe—as opposed to not-severe—case information.

In a similar study, Stevenage and Bennett (2017) presented laypeople with pairs of fingerprints that did or did not come from the same source. Each pair was accompanied by a statement summarising DNA test results, which indicated either (1) a match between the suspect and perpetrator, (2) no match between the suspect and perpetrator, or (3) insufficient evidence to allow a conclusion (control condition). Contextual information was manipulated in a within-subjects fashion. Signal detection analyses revealed that participants showed a liberal response bias—that is, they made a disproportionate number of match decisions relative to non-match decisions. Response bias was highest, however, when the contextual information suggested a DNA match. Participants' accuracy was also affected by the contextual information—relative to control trials, participants made more erroneous decisions when the DNA results were inconsistent with the ground truth, and more correct decisions when the DNA evidence was consistent.

The Present Study

Only two studies have directly investigated the effects of contextual information on handwriting analysis and comparison (Kukucka & Kassin, 2014; Miller, 1984), and only one unpublished study (reported in Merlino, 2015) looked at whether these effects also occur in signature examinations. Furthermore, none of these studies has considered their data using a signal detection framework. The aim of Study 1A was to use a signal detection paradigm to investigate the extent to which contextual information can influence performance on a signature-comparison task. To do this, we presented laypeople with a series of 20 trials that included either low-biasing or high-biasing contextual information about the case. On each trial, participants were required to compare a questioned signature with a set of four known signatures to determine whether the questioned signature was written by the same person as the known signatures (i.e., a genuine signature; signal absent), or someone other than the author of the known signatures (i.e., a forged signature; signal present). The questioned signatures were either genuine or forged, but the cues in the contextual information were directed towards a forgery. Considering that sensitivity did not change as a function of contextual information in Searston et al.'s study (2015), we hypothesized that participants' sensitivity would not differ between low- and high-bias trials, but that the presence of high-biasing contextual information would be associated with a more liberal response bias (an overall tendency to conclude that a signature was forged), relative to low-biasing contextual information.

STUDY 1A

Method

Participants

The participants were 81 undergraduate students from the University of Otago, New Zealand (51 females; *M* age = 20.2 years, *SD* = 3.03, range = 18 to 41 years). Analyses in G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) indicated that we had a sufficient sample size to detect a medium effect size with 80% power. Participants were recruited via the Psychology Participant Pool, which is a department-run system whereby 100- and 200-level psychology students can earn a small portion of course credit for completing a questionnaire after taking part in an experiment. All participants were informed of the purpose of the study and how the data would be used and stored; all gave their written consent to participate.

Design

We utilised a 2 x 2 within-subjects design, with Ground truth (genuine, forged) and Context (low-bias, high-bias) as the manipulated variables. There were two dependent variables: how certain participants were that the signature was forged prior to seeing the samples (as a manipulation check of the contextual information), and the participants' determination of whether each signature was forged or genuine.

Materials

Signatures.

We collected a sample of genuine signatures from 40 adult volunteers (*writers*) who were each provided with a sheet of A4 paper and three pens of varying thicknesses. Each writer was asked to write their signature as they would when signing an official document and to provide a total of nine signatures; three with each pen.

We subsequently asked a separate group of 10 adult volunteers (*forgers*) to create forgeries of these signatures. Each forger was presented with 10 randomly-selected signatures from the 40 writers, and asked to create as realistic a forgery as possible for each. The forgers were instructed not to trace a signature directly, as this approach results in features that can easily be identified as forgery (e.g., a perfect match when superimposed; Osborn, 1929). Forgers took as much time as they needed for each forgery, and could make as many attempts as they liked. Finally, we asked each forger to indicate which of their forgeries for each of the 10 signatures they would use in a real-life situation. We used those forgeries as our final stimuli.

A third set of adult raters (N = 31) judged the complexity of each genuine signature based on how easy they thought it would be to forge. Ratings were conducted on a 7-point scale, with 1 being "very easy" and 7 being "very difficult". The average rating for the 40 signatures was 3.58 (SD = 0.99, range = 5.37). We randomly selected 10 signatures for the forged condition, and then matched each to a genuine signature based on complexity rating. We then randomly selected four genuine samples for each of the resulting 20 signatures to serve as the known exemplars. A fifth was used in trials in which the questioned signature was genuine. In trials in which the questioned signature was forged, the questioned signature was a forgery (see Figure 6.2 for an example of a questioned signature accompanied by four known exemplars).

Figure 6.2. Example of a questioned signature accompanied by four known exemplars. Here, the questioned signature is genuine.

Contextual information.

The contextual information used in the experiment comprised short vignettes containing information relating to a signature comparison case. A different vignette accompanied each of the 20 signature sets. In the low-bias condition, the vignette contained information that simply indicated that the questioned signature was made by a suspect, or that police investigators sought to determine whether the signature was genuine or forged (e.g., “The police want to know whether the signature on a seized cheque was forged”). In the high-bias condition, the contextual information contained additional cues to suggest that the questioned signature was forged (e.g., “Police investigators want to know whether a suspect in a fraud case forged several signatures. The suspect initially confessed, but later retracted the confession stating that he had been coerced"). These additional cues supposedly came from one of the following sources: a retracted confession (as in the example given above), CCTV footage, an eyewitness statement, a pre-existing criminal record, or a fingerprint found on the questioned document. Each contextual information type was paired with two questioned signatures that were forged and two questioned signatures that were genuine.

Several issues related to the design[7] of our experiment, as well as software[8] limitations, resulted in us being unable to fully randomize signature and scenario pairings. Therefore, we manually controlled for possible confounding factors by ensuring that the signatures in the high-bias and low-bias conditions were of similar complexity. For example, if there was a forged signature with an average complexity rating of 4.83 in the high-bias condition, we assigned a forged signature with the closest complexity rating to the low-bias condition.

Procedure

Participants completed an initial practice trial, followed by 20 experimental trials presented in a random order. The data from the practice trial were not included in any subsequent analyses. On each trial, participants were first asked to read the contextual information and rate on a 100-point Likert scale (0 = certain not forged, 100 = certain forged) how certain they were that the questioned signature had been forged. We used these ratings as a manipulation check for our contextual information manipulation. Next, participants were presented with a questioned signature alongside four known signatures. Instructions presented on the computer screen asked participants to carefully examine the questioned signature and known signatures to determine, in their own time, if the questioned signature was forged (i.e., written by someone other than the person who wrote the known signatures)

[7] Each trial consisted of three separate components, which none of the software available to us could reliably randomize. Additionally, many of our case scenarios included names visible in the signatures to make them appear more realistic. This aspect of the design meant that we could not assign signatures to scenarios randomly.
[8] We presented the stimuli on *Qualtrics*, a web-based survey tool.

or genuine (i.e., written by the same person who wrote the known signatures). Consistent with our signal detection approach, we only provided participants with two response options: forged or genuine. When participants had completed all 20 trials, they were thanked and fully debriefed.

Results

Manipulation Check

After presenting participants with the contextual information but before exposing them to the signatures, we asked them to rate their level of certainty that the signature was forged on a scale of 0 to 100. We did this to ensure that our high-bias trials were more likely to create the expectation of a forgery than our low-bias. As expected, a paired-samples *t*-test on these data revealed that participants were more likely to expect that the signature would be forged on high-bias trials (M = 71.21, SD = 9.77), than on low-bias trials (M = 52.84, SD = 9.03), $t(80)$ = 15.33, p < .001, Cohen's d = 1.70, 95% CI [15.99, 20.76].

Signal Detection Analysis

Please note that in our experiment, the signal is differences (i.e., signs of a forgery). For each participant, we calculated the number of hits, correct rejections, false alarms, and misses, separately for the high-bias and low-bias trials. These values are presented in Table 6.1. We used these values to calculate sensitivity (d') and response criterion (c).

Table 6.1.

Mean Number (SD) of Hits, Misses, False Alarms, and Correct Rejections on Low-bias and High-bias Trials.

	Hit	Miss	False Alarm	Correct Rejection
Low-bias	4.22 (0.99)	0.81 (0.87)	1.56 (0.96)	3.40 (1.05)
High-bias	4.23 (0.93)	0.78 (0.92)	1.52 (1.07)	3.47 (1.08)

Sensitivity. On the high-bias trials (M = 1.39, SD = 0.74), *d'* differed significantly from 0, $t(80) = 16.91$, $p < .001$, $d = 1.88$, 95% CI [1.22, 1.55], as did *d'* on the low-bias trials (M = 1.31, SD = 0.66), $t(80) = 17.77$, $p < .001$, $d = 1.97$, 95% CI [1.16, 1.46]. Furthermore, a paired-samples *t*-test revealed that *d'* did not differ significantly as a function of the contextual information manipulation, $t(80) = 0.84$, $p = .40$, $d = 0.09$, 95% CI [-0.12, 0.27]. That is, participants were able to discriminate between forged and genuine signatures at a rate higher than chance, and their level of discrimination did not differ as a function of context (see Figure 6.3).

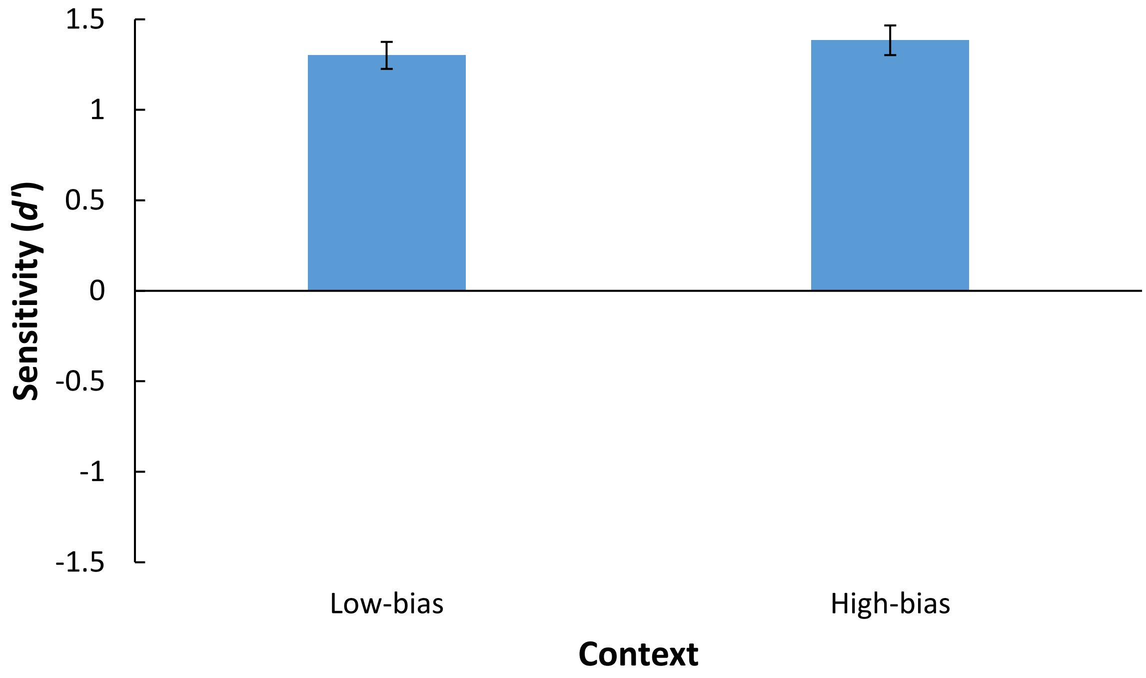


Figure 6.3. Mean sensitivity (*d'*; $\pm$1*SE*), shown as a function of Context (low-bias, high-bias).

Response criterion. Next, we examined participants' response criterion (*c*) values. Note that *c* values less than zero reflect a bias towards concluding that the signature is genuine, while *c* values greater than zero reflect a bias towards concluding that the signature is forged. One-sample *t*-tests revealed that criterion location on both high-bias trials, $t(80)$ =

4.53, $p < .001$, $d = 0.50$, 95% CI [0.11, 0.29], and low-bias trials, $t(80) = 4.82$, $p < .001$, $d = 0.54$, 95% CI [0.13, 0.30], differed significantly from 0. Participants displayed a tendency to state that a signature was forged, both on high-bias ($M = 0.20$, $SD = 0.40$), and low-bias ($M = 0.21$, SD, 0.39) trials. Response criterion did not differ significantly as a function of Context, $t(80) = -0.17$, $p = .87$, $d = -0.02$, 95% CI [-0.10, 0.08] (see Figure 6.4).

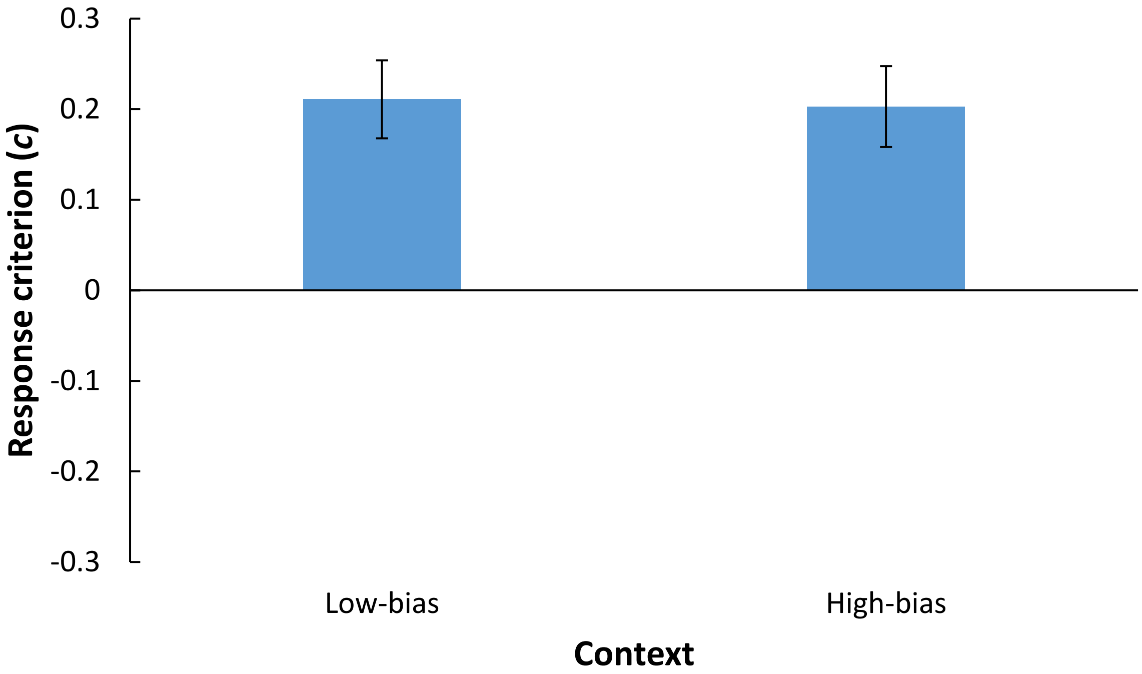


Figure 6.4. Mean response criterion (c; $\pm 1SE$), shown as a function of Context (low-bias, high-bias).

Accuracy. Participants' accuracy is a function of both sensitivity and response criterion. As such, we calculated the percentage of correct decisions as a function of ground truth, averaged separately for low-bias and high-bias trials. A paired-samples t-test revealed that participants made a significantly higher percentage of correct responses when the questioned signature was forged (84.57%) than when it was genuine (68.76%), $t(80) = 5.35$, $p < .001$, $d = 0.96$, 95% CI [11.55, 20.06]. We also conducted paired-samples t-tests to determine whether accuracy on genuine and forged signatures was affected by context (see Table 6.2). We did not find any significant differences in accuracy between low-bias and high-bias trials, both when the questioned signature was genuine and when it was forged.

Table 6.2.

Accuracy as a Function of Ground Truth and Context.

Ground truth	Context	Mean (%)	*SD*	*t*	*p*
Genuine	Low-bias	68.15	20.92	-	
				0.45	.66
	High-bias	69.38	21.70	-	
Forged	Low-bias	84.44	19.75	-	
				0.10	.92
	High-bias	84.69	18.51	-	

We also investigated the effect of contextual information on our participants' final decision. We calculated the average number of forged decisions on high-bias and low-bias trials for each participant, regardless of ground-truth. A paired-samples *t*-test to compare the mean number of forged decisions on low-bias (M = 5.72, SD = 1.49) and high-bias (M = 5.77, SD = 1.42) trials resulted in no significant difference, $t(80) = 0.27$, $p = .79$, $d = 0.02$, 95% CI [-0.33, 0.38].

Given that we found no effect of context despite successful manipulation checks, we wanted to investigate whether participants' expectation of forgery (i.e., certainty rating in the manipulation check) correlated with the examination outcome. We averaged the certainty ratings separately for high-bias and low-bias trials for each participant and compared those to the mean number of forged decisions on high-bias and low-bias trials. Certainty ratings were not significantly correlated with examination outcome on both low-bias (r = -.05) and high-bias trials (r = .10).

STUDY 1B

In Study 1A, we did not find an effect of contextual information on participants' sensitivity or response criterion. Although our manipulation check confirmed that high-biasing case information increased participants' expectations that the signature would be a forgery, these expectations did not translate to participants' decisions about the signatures. Although the most conservative interpretation of these data is that contextual information does not exert a significant effect on signature comparisons, there are at least two other explanations that warrant further investigation.

One possible explanation for the lack of context effect, for example, is that the contextual information was not tied well enough to the decision-making process. This could explain why contextual information affected participants' expectations of a forgery, but not their final judgments. One way to address this possibility would be to present the contextual information at the same time as the signatures, rather than directly beforehand.

A related possibility is that even though the difference between low-bias and high-bias information was enough to bias expectations, it was not strong enough to impact decision-making. We wondered, for example, whether the way the manipulation check question was worded set up some expectation that the signature was forged, regardless of whether the trial was low-bias or high-bias. A considerable body of research has shown that wording of a question can influence not only the answer given (e.g., Loftus & Palmer, 1974; Loftus & Zanni, 1975), but also the response to subsequent questions (Loftus & Palmer, 1974). For example, Loftus and Palmer (1974) presented participants with a video of a car accident and asked them how fast the cars had been going prior to the accident. The experimenters manipulated the verb used to describe the accident; some participants were asked how fast the cars were going when they *hit* each other, and some were asked how fast the cars were going when they *smashed into* each other. When the participants were later asked if they had seen any broken glass at the scene, those in the *smashed into* condition were twice as likely to report seeing broken glass compared to those in the *hit* condition. In Study 1A, our manipulation check only asked participants to indicate their certainty that the questioned signature was forged. Using the word *forged* in the manipulation check could have created an anchoring bias in our participants, in which any subsequent considerations of the signature's authenticity would be thought of in terms of forgery (Tversky & Kahneman, 1974).

It is also possible that our dichotomous response scale was simply not sensitive enough to detect bias—that is, that the dichotomous nature of our task obscured context effects. While the advantage of a signal detection framework is that it allows us to tease apart sensitivity and response bias, it does not take into account that people might hold varying levels of confidence in the same decision. More detailed measures can provide other insights into the nuances of decision-making. Kukucka and Kassin (2014), for example, observed that context influenced participants' perceptions of the questioned handwriting, which resulted in an increased belief in the suspects' guilt.

We addressed each of these possibilities in Study 1B. Specifically, we simplified and changed the placement of the contextual information. We also changed our dependent variable from a dichotomous scale to a 12-point continuous scale. Although this scale still required participants to decide whether or not each signature was forged, it also allowed us to examine the effect of context on the confidence with which they held their decisions.

Method

Participants

The participants were 85 undergraduate students from the University of Otago, New Zealand (70 females; *M* age = 24 years, *SD* = 6.85, range = 18 to 59 years). Analyses in G*Power (Faul et al., 2009) indicated that we had a sufficient sample size to detect a medium effect size with 80% power. Participants were recruited via the Psychology Participant Pool, which is a department-run system whereby 100- and 200-level psychology students can earn a small portion of course credit for completing a questionnaire after taking part in an experiment. All participants were informed of the purpose of the study and how the data would be used and stored; all gave their written consent to participate.

Design

We utilised a 2 x 2 within-subjects design, with Ground truth (genuine, forged) and Context (low-bias, high-bias) as the manipulated variables. There was one dependent variable: a forced-choice confidence rating regarding the questioned signature's authenticity.

Materials

We used the same signatures as in Study 1A, but we simplified the contextual information so that it could be presented alongside the questioned and known signatures. Instead of presenting participants with a vignette of case-related information, we presented participants with a simple statement that told the participant what other forensic experts had concluded about the signature. On low-bias trials the contextual information always stated that *"50% of forensic experts think that the questioned signature below has been forged"*, implying that the other 50% of experts thought it was genuine. In the high-biasing condition, the percentage agreement of forensic experts ranged from 60% to 100%.

Procedure

As in Study 1A, participants completed an initial practice trial, followed by 20 experimental trials presented in a random order. The data from the practice trial were not included in any subsequent analyses. In each trial, participants saw one questioned signature and four known exemplars. We asked participants to examine the genuine and questioned signatures carefully, and to indicate their confidence that the questioned signature had been forged or was genuine. We recorded their responses on a forced-choice confidence scale, which ranged from 1 (sure genuine) to 6 (guess genuine) and 7 (guess forged) to 12 (sure forged). When participants had completed all 20 trials, they were thanked and fully debriefed.

Results

Signal Detection Analysis

As in Study 1A, we calculated the number of hits, misses, false alarms, and correct rejections for each participant; we did this separately for the low-bias and high-bias trials (see Table 6.3 for a summary). Please note that in our experiment the signal is differences (i.e., signs of a forgery). Participants gave their responses on a forced choice confidence scale, which ranged from 1 (*sure genuine*) to 12 (*sure forged*). For the signal detection analysis, ratings of 1 to 6 were considered as a 'genuine' response, and ratings of 7 to 12 were considered as a 'forged' response. We used the response tallies to calculate sensitivity (*d'*) and response criterion (*c*).

Table 6.3.

Mean Number (SD) of Hits, Misses, False Alarms, and Correct Rejections on Low-bias and High-bias Trials.

	Hit	Miss	False Alarm	Correct Rejection
Low-bias	3.80 (1.08)	1.2 (1.08)	1.96 (1.17)	3.04 (1.17)
High-bias	3.94 (0.82)	1.06 (0.92)	3.51 (1.10)	1.45 (1.07)

Sensitivity. As in Study 1A, sensitivity (*d'*) served as our measure for the participants' ability to distinguish genuine from forged signatures. A *d*' value of 0 indicates that participants could not discriminate between forged and genuine signatures at a rate above chance. On low-bias trials (M = 0.90, SD = 0.71), *d'* differed significantly from zero, $t(84)$ = 11.59, p < .001, Cohen's d = 1.26, 95% CI [0.07, 0.37], as did *d'* on high-bias trials, (M = 0.22, SD = 0.70), $t(84)$ = 2.86, p < .01, d = 0.31, 95% CI [0.74, 1.05]. That is, regardless of context, participants were able to discriminate genuine from forged signatures at a rate greater than chance. A paired-samples *t*-test revealed that sensitivity was significantly lower on high-bias trials (M = 0.22, SD = 0.70) than on low-bias trials (M = 0.90, SD = 0.71), $t(84) = 6.89$, $p < .001$, $d = 0.75$, 95% CI [0.48, 0.87] (see Figure 6.5).

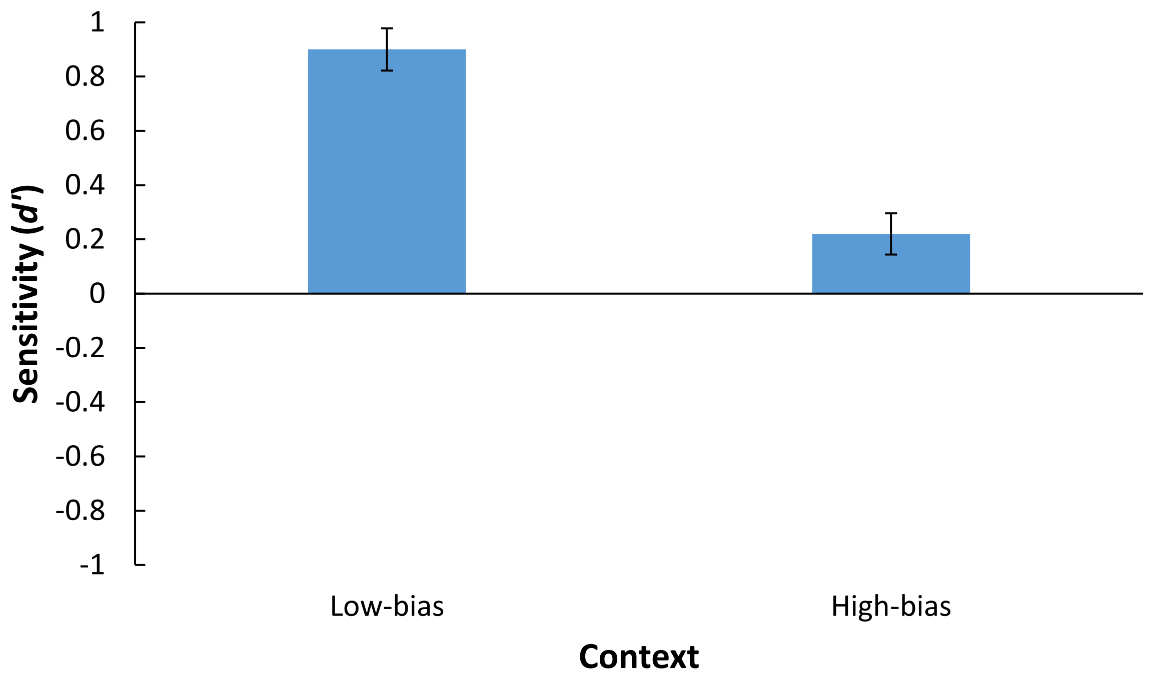


Figure 6.5. Mean sensitivity ($\pm$1*SE*) as a function of Context (low-bias, high-bias).

Response criterion. Next, we examined criterion location (c) to investigate response bias; c values less than zero reflect a tendency to conclude that a questioned signature is genuine, while c values greater than zero reflect a tendency to conclude that a questioned signature is forged. Criterion location on low-bias trials (M = .21, SD = .46) differed significantly from zero, $t(84) = 4.15$, $p < .001$, $d = 0.45$, 95% CI [.11, .31], as did criterion location on high-bias trials (M = .62, SD = .41), $t(84) = 14.02$, $p < .001$, $d = 1.52$, 95% CI [.53, .71]. A paired-samples *t*-test revealed that response criterion was significantly higher on high-bias trials (M = .61, SD = .41) than on low-bias trials (M = .21, SD = .46), $t(84) = 6.81$, $p < .001$, $d = 0.74$, 95% CI [.29, .53] (see Figure 6.6). That is, regardless of context, participants showed a tendency to conclude that the signatures were forged, but this tendency was amplified on high-bias trials.

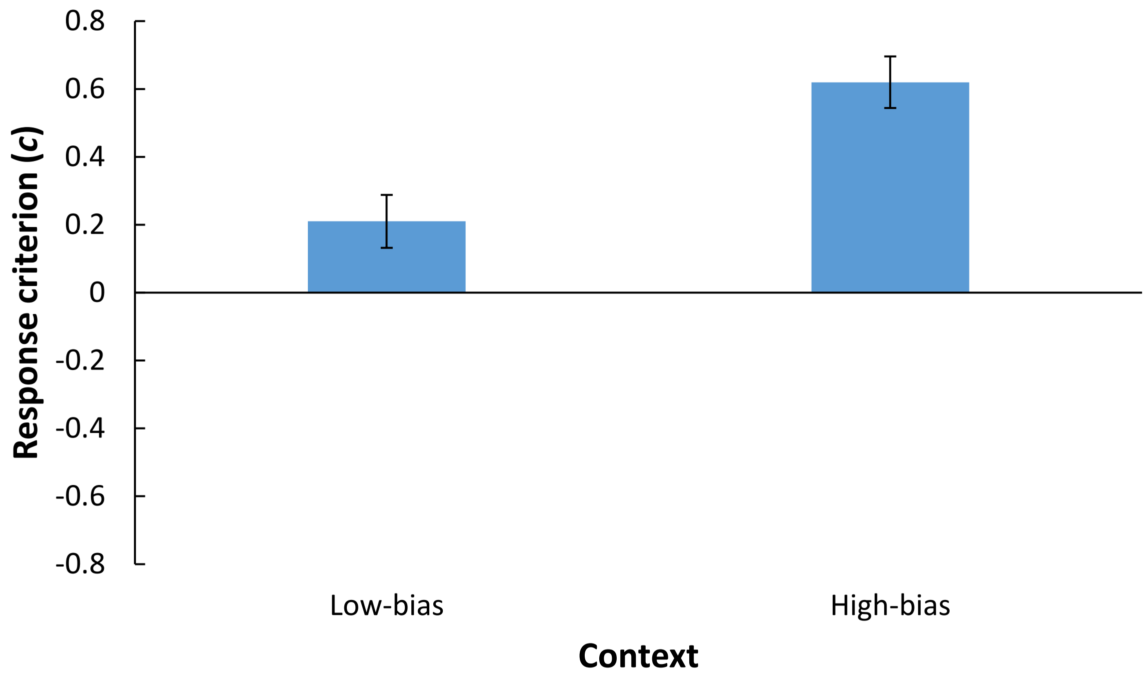


Figure 6.6. Mean response criterion (±1*SE*) as a function of Context (low-bias, high-bias).

Confidence Ratings

We asked participants to report their decision using a forced-choice 12 -point confidence scale (1 = sure genuine; 6 = guess genuine; 7 = guess forged; 12 = sure forged). To determine the strength of participants' confidence in their decisions, we recalculated their ratings on this scale as a function of their decision to reflect that a rating of 1 demonstrates same degree of confidence as a rating of 12. As such, for subsequent analyses, confidence ratings range from 1 to 6, with 6 being the highest degree of confidence and 1 being the lowest.

We conducted a repeated-measures ANOVA with Context (low-bias, high-bias) and Decision (genuine, forged) as the within-subjects variables. Significant main effects of Context, $F(1, 84) = 9.96$, $p < .01$, $\eta_p^2 = .11$, and Decision, $F(1, 84) = 19.89$, $p < .001$, $\eta_p^2 =$

.19, were qualified by a significant Context x Decision interaction, $F(1, 84) = 25.08$, $p < .001$, $\eta_p^2 = .23$. Context did not affect participants' confidence in their 'forged' decisions, paired-samples $t(84) = 1.63$, $p = 0.11$, $d = -0.17$, 95% CI [-0.34, 0.03], (low-bias $M = 3.11$, $SD = 0.99$; high-bias $M = 3.26$, $SD = 0.93$) . It did, however, affect their confidence in their 'genuine' decisions, paired-samples $t(84) = 5.41$, $p < .001$, $d = 0.56$, 95% CI [0.39, 0.86], (low-bias $M = 3.06$, $SD = 0.96$; high-bias $M = 2.47$, $SD = 0.98$). Specifically, confidence in 'genuine' decisions was significantly reduced on high-bias trials, relative to low-bias trials (see Figure 6.7).

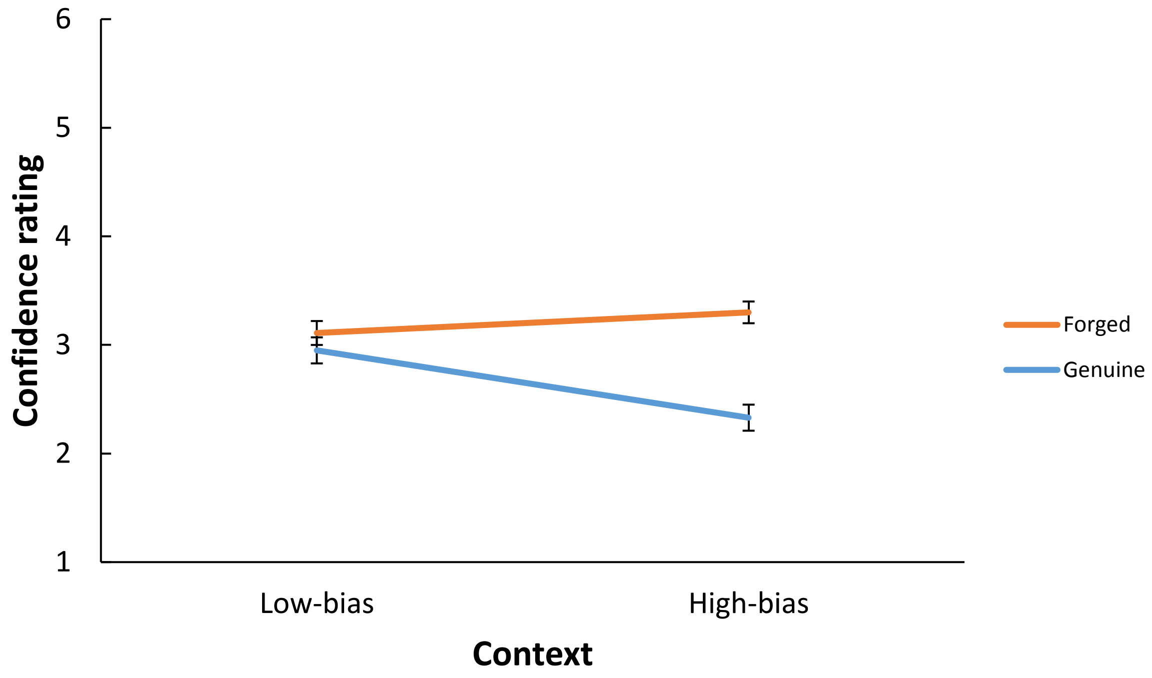


Figure 6.7. Mean confidence ratings ($\pm 1SE$), shown as a function of Context (low-bias, high-bias) and Decision (genuine, forged).

Discussion

Previous studies on contextual bias in forensic examiners have typically been restricted to comparing match or non-match responses made in the presence or absence of context (e.g., Dror et al., 2005; Kukucka & Kassin, 2014; Miller, 1984; Osborne & Zajac, 2015), or measuring intra-examiner consistency after exposure to potentially biasing information (e.g., Dror & Charlton, 2006; Dror & Hampikian, 2011). In two experiments, we examined the effect of contextual information in a signature comparison task utilising a signal detection approach. Such an approach allowed us to take a closer look at how biasing information might exert its effect.

Contrary to our expectations, in Study 1A we did not observe an effect of context manipulation on participants' decisions. Although our manipulation check confirmed that high-biasing case information led to significantly higher expectations of forgery than low-biasing case information, these expectations did not translate to participants' decisions about the signatures. That is, we did not find a significant correlation between participants' prior expectations of forgery and the subsequent examination outcome. Neither did we observe significant effects of our context manipulation on sensitivity (*d'*) or response criterion (*c*). In other words, the contextual information did not exert a significant influence on participants' ability to distinguish forged signatures from genuine ones, or their tendency to decide that the signature was forged.

In Study 1B, we modified our procedure to address several potential reasons why we did not detect a context effect. Specifically, we more rigidly tied the contextual information to the signatures, by presenting it to participants at the same time as the test stimuli, rather than beforehand. We also asked participants to make their decisions on a 12-point forced choice confidence scale, rather than making decisions that were restricted to 'genuine' or 'forged.' This second modification allowed us to examine the data in two ways, exploring the possibility that the dichotomous response option was not sensitive enough to detect any context effects.

Contrary to Study 1A, in Study 1B we observed significant effects of context manipulation on both sensitivity (*d'*) and response criterion (*c*)[9]. Highly-biasing contextual information reduced our participants' ability to distinguish forged signatures from genuine ones, and lowered their threshold for deciding that questioned signatures were forged. When we considered participants' confidence in their decisions, we also saw a significant effect of context. Although context did not exert a significant effect on participants' confidence in their 'forged' decisions, it reduced the degree to which they were confident in their 'genuine' decisions. These findings are similar to those of Kukucka and Kassin (2014), who discovered a significant correlation between their participants' match-confidence and guilt-confidence scores. Participants who were less confident that the suspect's handwriting matched the questioned writing were also less confident that the suspect was guilty, especially when informed of the suspect's retracted confession.

Notably, regardless of the contextual information manipulation, participants in both experiments displayed an overall tendency—as measured by response criterion—to conclude that the questioned signature was forged. It is possible that the task itself might have set up such a response bias. While removing the manipulation check used in Study 1A reduced the risk of an anchoring bias, the mere fact that the authenticity of the signatures was in question may have led the participants to anticipate that they were forged. Lange et al. (2011), for example, showed that simply using the word 'suspect' created context-driven expectations, whereby participants became biased towards interpreting an ambiguous audio statement as containing incriminating statements. As discussed in Chapter 2, an individual's expectations and beliefs directly influence what information they attend to, how they interpret and subsequently remember the information, as well as any decisions made based on that

[9] Please note, however, that because we changed multiple variables between Study 1A and 1B we cannot attribute the difference in results to any one of these.

information (Charman, Gregory, & Carlucci, 2009). Context-driven expectations of guilt have been shown to affect police interviewers (Hill, Memon, & McGeorge, 2008; Narchet, Meissner, & Russano, 2011), jurors (Devine, Clayton, Dunford, Seying, & Pryce, 2001), judges (Halverson, Hallahan, Hart, & Rosenthal, 1997), and eyewitnesses (Hasel & Kassin, 2009).

In Studies 1A and 1B, examining a suspicious signature may have been enough to bias the participants towards declaring a forgery because it fit with their expectations based on the task they were given. The inclusion of a control group with no contextual information (i.e., simply presenting participants with signatures in the absence of any information about what other forensic examiners concluded) would allow us to test this hypothesis. Findings from other studies, however, suggest that a control group would still have demonstrated biased decision-making. Tangen, Thompson, and McCarthy (2011) found that laypeople tend to have a liberal response bias and tend to declare two fingerprints a match even in the absence of any contextual information. Likewise, the participants in Searston et al. (2015) displayed a liberal response bias regardless of whether the contextual information was present or not. Stevenage and Bennett (2017) demonstrated similar findings; participants showed a significant liberal bias on control trials, where the participants determined two fingerprints to match even in the absence of biasing information (see Chapter 9, for further discussion of this issue).

It is important to note that we did not give our participants the option of making an inconclusive decision—a response that would always be available to a forensic examiner. Doing so would have prevented us from analysing the data using a signal detection approach. Although this approach was appropriate for our current study, it also resulted in us being limited in comparing our findings to those of studies that gave their participants an inconclusive option (e.g., Dror & Hampikian, 2011; Miller, 1984; Osborne & Zajac, 2015;

Smalarz, Madon, Yang, Guyll, & Buck, 2016). Future research should endeavour to develop signal detection paradigms that nevertheless allow for the inclusion of an inconclusive option (see Chapter 9 for further discussion of this issue).

Where to from here?

Few studies have investigated the potential for contextual bias in the forensic examination of handwriting (Kukucka & Kassin, 2014; Merlino, 2015; Miller, 1984), and none have teased apart the roles of sensitivity and response bias. We found mixed evidence to suggest that these variables—as well as confidence—can be influenced by contextual information, in laypeople at least (see Chapter 9, for a discussion of possible differences between experts and novices). Interestingly, our data from Study 1B suggest that contextual information might affect the signature examination *process*—a possibility that has yet to be investigated in the empirical literature (Dror, 2009; Kukucka, 2014). That is, contextual information may change a person's search for features in the evidence, rather than just their evaluations of those features. In Study 1B, the effect of context was not limited to a shift in response criterion; high-biasing information also reduced participants' ability to discriminate between the signatures. One possible explanation for this finding is that directional cues in the contextual information alter the way in which people compare the signatures—perhaps by promoting the use of a positive test strategy (Ask & Granhag, 2005; Klayman & Ha, 1987). Information suggesting that a signature has been forged, for example, might encourage people to focus primarily—or even exclusively—on differences between the questioned and known signatures, while dismissing the similarities. We investigated this possibility more closely in Study 2.

CHAPTER SEVEN

STUDY 2: CAN THE WAY IN WHICH LAYPEOPLE EVALUATE QUESTIONED SIGNATURES CHANGE THE EFFECT OF CONTEXTUAL INFORMATION?

In Study 1, we found some evidence that directional cues present in contextual information can influence people's decisions about whether a signature is genuine or forged. But these findings tell us little about why these effects emerge. The presence of a context effect on sensitivity in Study 1B suggests that, rather than simply affecting participants' decision threshold, the contextual information could have influenced the *process* by which participants examined the signatures—a possibility that has not been directly investigated in the literature (Dror, 2009; Kukucka, 2014). In Study 2, we therefore focused on the way in which participants compared the signatures.

Confirmation Bias in the Examination Process

The key component of the handwriting examination process is a side-by-side comparison of the questioned and known writing. As discussed in Chapter 4, it is crucial that the examiner thoroughly evaluates both the differences and the similarities in the material under consideration. As Osborn (1929), the founder of modern document examination, stated:

> The process is always a double operation, positive and negative, and if error is to be avoided neither part of the process should be overlooked. In order to reach the conclusion of identity of two sets of writings there must not be present significant and unexplained divergences. (p. 262)

When examining questioned and known signatures, examiners carry out a visual search of the evidence to identify and evaluate similarities, differences, and unique features. Exposure to case information—for example, that the suspect has made a confession—could unduly influence the examiner's search for and evaluation of these features (Sulner, 2014). More specifically, expectations of a certain outcome can cause disproportionate attention to features in the evidence that confirm those expectations (Risinger et al., 2002).

According to Sulner (2014), such expectancies might elicit a *positive test strategy*, in which an examiner focuses on features in the writing that support the expected outcome. If, for example, contextual information suggests that the questioned and known writing share the same source, the expert's attention might be disproportionately focused on looking for similarities in the writing and fail to adequately search for, or recognize, dissimilarities (Risinger et al., 2002; Sulner, 2014).

The idea of a positive test strategy was first proposed by Klayman and Ha (1987), and describes an approach to hypothesis testing that involves choosing to examine instances in which the expected feature or event is expected to occur, or choosing to examine instances in which the feature of interest is known to have occurred. For example, Wason (1960) presented participants with sets of three numbers (*triples*) that conformed to a rule, and asked them to propose new triples that would help them discover what the rule was (see Chapter 3). In Wason's experiments, participants were much more likely to propose triples that would confirm—rather than disconfirm—their hypothesis. That is, they were using a positive test strategy (Klayman & Ha, 1987).

People's tendency to seek information that is consistent with their prior beliefs and expectations has been demonstrated in a variety of rule-discovery tasks (Klayman & Ha, 1987; Mynatt, Doherty, & Tweney, 1978; Tweney et al., 1980), in various domains such as social psychology (Snyder, 1981; Snyder & Swann, 1978), and decision-making (Fischer, Jonas, Frey, & Schulz-Hardt, 2005; Kray & Galinksy, 2003; Schulz-Hardt, Frey, Lüthgens, & Moscovici, 2000). In general, the research suggests that poor decision-making can often be attributed to the tendency to default to a positive test strategy, when seeking disconfirming information (i.e., using a negative test strategy) would result in better decisions (Kray & Galinsky, 2003).

The Present Study

The aim of Study 2 was to investigate the association between the effect of contextual information and the signature examination process. Due to the difficulties inherent in measuring the processes that participants use in their decision-making, we took a different approach: we *manipulated* the process that participants were instructed to use. That is, some participants were instructed to identify differences between the signatures, while others were instructed to identify similarities.

If exposure to high-biasing contextual information promoted the use of a positive test strategy in our Study 1B participants, then it should be possible to eliminate contextual bias by encouraging them to use a negative test strategy instead (i.e., to focus on similarities). Therefore, we expected that participants in our differences condition would show lower sensitivity and a higher response criterion in high-bias trials, relative to low-bias trials. We also expected participants in the differences condition to identify more points when exposed to high-biasing information than when exposed to low-biasing information. We did not expect to see any of these effects in the similarities condition.

Method

Participants

The participants were recruited via two methods. Some participants were undergraduate psychology students from the University of Otago, New Zealand, who received a small portion of course credit for completing a questionnaire at the end of the experience. Others were community members who signed up for the experiment via the Psychology's Research Participation website. These participants were compensated for costs incurred in participating. The final sample comprised 186 participants (50 males, 136 females; *M* age = 22 years, *SD* = 4.42, range = 18 to 47 years). Analyses in G*Power (Faul et al., 2009) indicated that we had a sufficient sample size to detect a medium effect size with 80% power. All participants were informed of the purpose of the study and how the data would be used and stored; all gave their written consent to participate.

Design

We utilised a 2 x 2 x 2 factorial design. There were two within-subjects factors: Ground truth (genuine, forged) and Context (low-bias, high-bias). There was one between-subjects factor: Task (similarities, differences); participants were quasi-randomly assigned to this factor, such that there were approximately equal numbers in each of the two conditions. There were two dependent variables: the number of points selected on the questioned signature, and a forced-choice confidence rating regarding the questioned signature's authenticity.

Materials

Signatures.

We used the same signatures as in Study 1. However, we reversed which signatures were assigned to the 'genuine' condition and which were assigned to the 'forged' condition. We also switched which signatures were assigned to the low-bias and high-bias conditions. That is, if a signature was assigned to the high-bias condition in Study 1, then it was assigned to the low-bias condition in Study 2. This was done to in an attempt to control for any possible confounding factors in the stimuli that we might not have been aware of.

Contextual information.

We used the same contextual information as Study 1B. On low-bias trials, the contextual information always stated that *"50% of forensic experts think that the questioned signature below has been forged"*, implying that remaining 50% considered the questioned signature was genuine. In the high-biasing condition the percentage of agreement from forensic experts ranged from 60% to 100% and was counterbalanced across ground truth.

Heat mapping software.

We used heat-mapping software offered by Qualtrics to quantify the points of similarity and difference identified by our participants. This software allowed the participants to indicate points of difference and similarity by clicking directly on the questioned signature using the computer mouse (see Figure 7.1).

Procedure

Participants completed an initial practice trial, followed by 20 experimental trials presented in a random order. The data from the practice trial were not included in analyses. As in Study 1, in each trial, participants saw one questioned signature and four known exemplars. We asked participants to examine the questioned and known signatures carefully. Approximately half of the participants were asked to select points on the questioned signature that they thought were significantly similar to the known signatures (*similarities* condition; $n = 92$); remaining participants were asked to select differences (*differences* condition; $n = 94$; see Figure 7.1 for a trial example). Regardless of task condition, participants then indicated their decision on a 12-point forced-choice confidence scale, which ranged from 1 (sure genuine) to 6 (guess genuine) and 7 (guess forged) to 12 (sure forged). When participants had completed all 20 trials, they were thanked and fully debriefed.

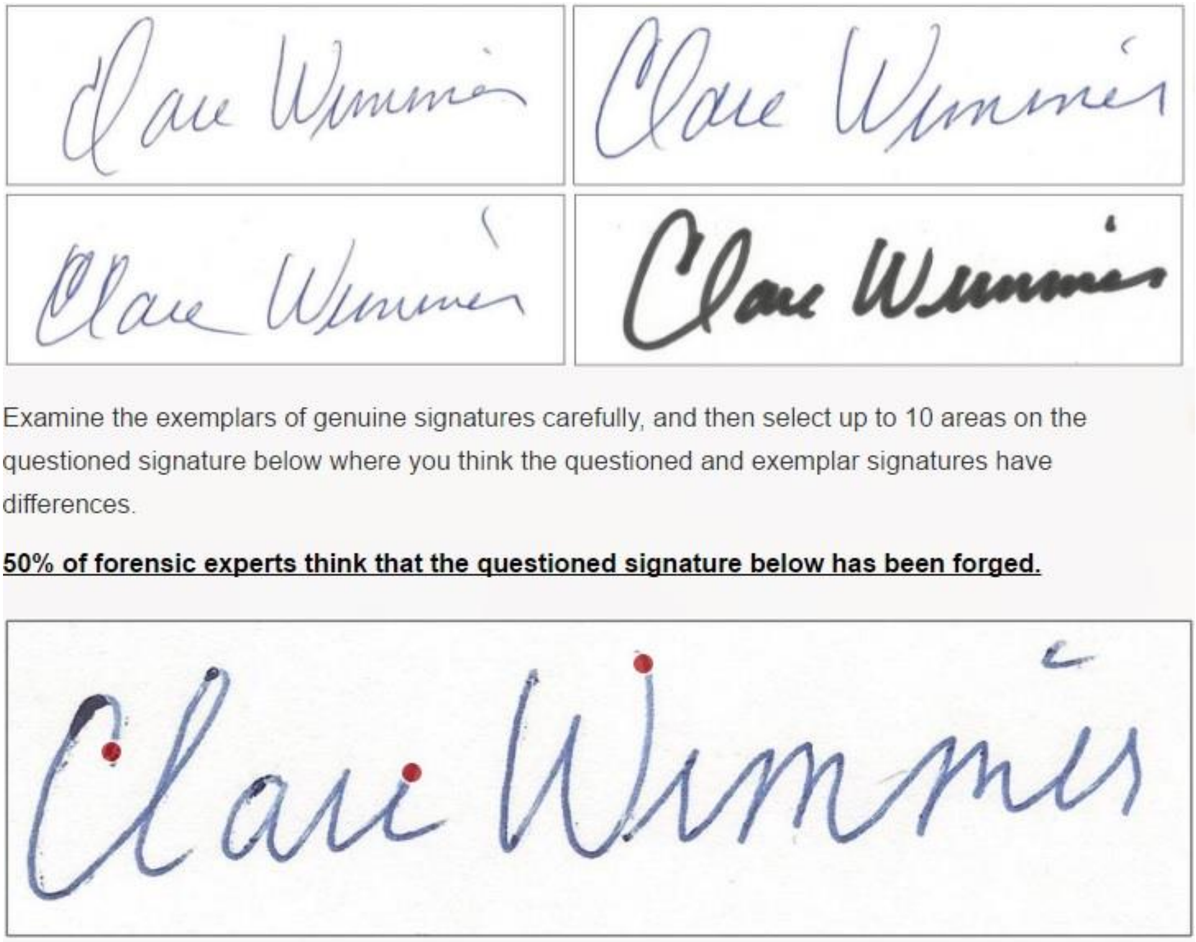


Figure 7.1. Example of a low-bias trial. The participant has selected three points of difference.

Results

Signal Detection Analysis

As in Study 1, we calculated the number of hits, misses, false alarms, and correct rejections for each participant; we did this separately for the low-bias and high-bias trials (see Table 7.1 for a summary). Please note that in our experiment the signal is differences (i.e., signs of a forgery). Participants gave their responses on a 12-point forced choice confidence scale, which ranged from 1 (sure genuine) to 6 (guess genuine), and 7 (guess forged) to 12 (sure forged). For the signal detection analysis, ratings of 1 to 6 were considered as a ‘genuine’ response, and ratings of 7 to 12 were considered as a ‘forged’ response. We used the response tallies to calculate sensitivity (*d’*) and response criterion (*c*).

Table 7.1.

Mean Number of Hits, Misses, False Alarms (FA), and Correct Rejections (CR) as a Function of Task and Context.

	Low-bias				High-bias			
	Hit	Miss	FA	CR	Hit	Miss	FA	CR
Similarities	3.83	1.17	1.95	3.05	4.01	0.99	3.49	1.51
Differences	3.23	1.77	1.76	3.24	3.67	1.33	3.41	1.59

Sensitivity. As in Study 1, sensitivity (*d'*) served as our measure for the participants' ability to distinguish genuine signatures from forged ones. A *d*' value of 0 indicates that participants could not discriminate between forged and genuine signatures at a rate above chance. We conducted an ANOVA with Task as the between-subjects variable and Context as the within-subjects variable (see Figure 7.2). There was a significant main effect of Context, $F(1, 184) = 89.25$, $p < .001$, $\eta_p^2 = .33$; sensitivity was higher in low-bias trials (M = 0.82, SD = 0.72) than in high-bias trials (M = 0.20, SD = 0.70). There was also a significant main effect of Task, $F(1, 184) = 5.18$, $p < .05$, $\eta_p^2 = .03$; participants in the similarities condition had, on average, higher sensitivity (M = 0.60, SD = 0.55) than participants in the differences condition (M = 0.42, SD = 0.54). There was no significant Context x Task interaction, $F(1, 184) = 0.20$, $p = .66$, $\eta_p^2 = .00$.

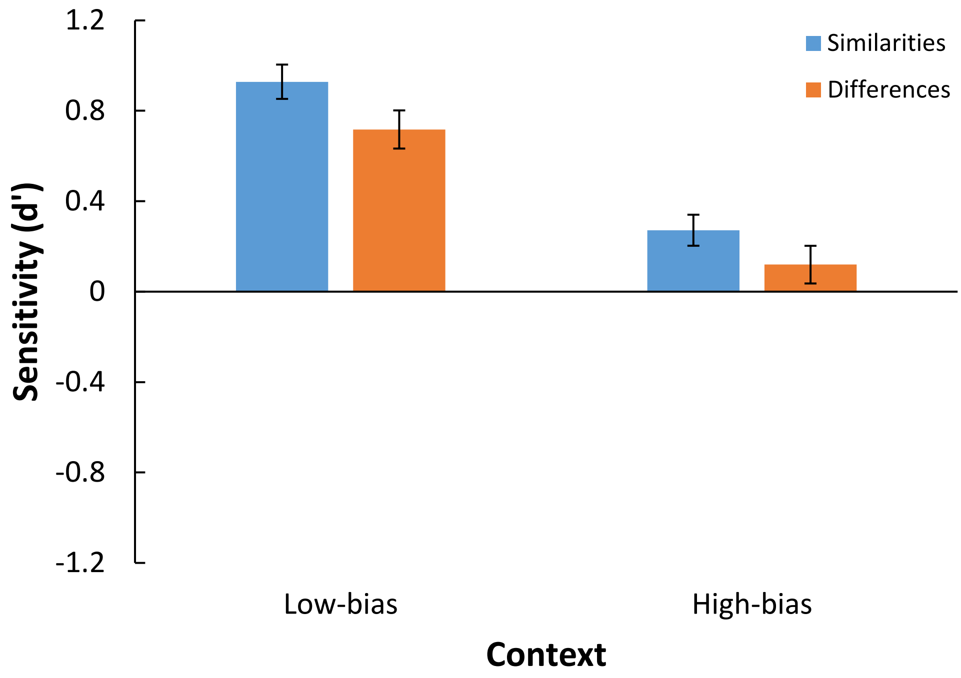


Figure 7.2. Mean sensitivity (*d'*; $\pm$1*SE*), shown as a function of Task (similarities, differences) and Context (low-bias, high-bias).

Response criterion. Next, we examined response criterion (*c*) to investigate response bias; *c* values less than zero reflect a bias towards concluding that the signature is genuine, while *c* values greater than zero reflect a bias towards concluding that the signature is forged. We conducted an ANOVA with Task as the between-subjects factor and Context as the within-subjects factor (see Figure 7.3). There was a significant main effect of Context, $F(1, 184) = 144.46$, $p < .001$, $\eta_p^2 = .44$; participants exhibited higher *c* values in high-bias trials ($M = .58$, $SD = .45$) than in low-bias trials ($M = .10$, $SD = .46$). That is, participants were more likely to conclude that the questioned signature was forged in high-bias trials than in low-bias trials. Furthermore, there was a significant main effect of Task, $F(1, 184) = 8.53$, $p < .01$, $\eta_p^2 = .04$; participants in the similarities ($M = .42$, $SD = .37$) condition exhibited, on average, higher *c* values than participants in the differences condition ($M = .26$, $SD = .37$). There was no significant Context x Task interaction, $F(1, 184) = 1.01$, $p = .32$, $\eta_p^2 = .01$.

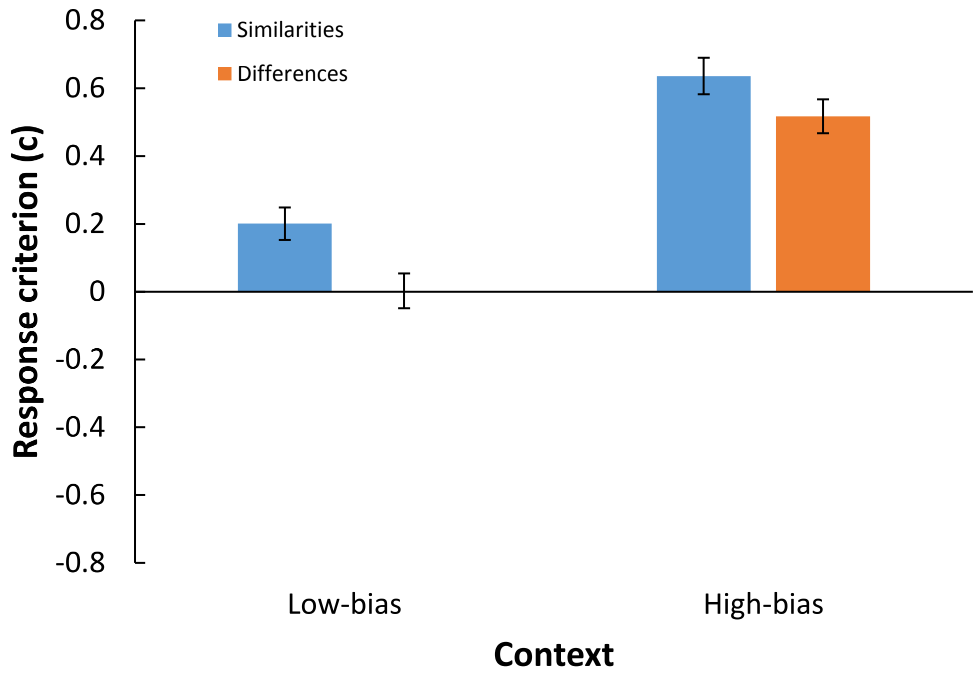


Figure 7.3. Mean response criterion (*c*; ±1*SE*), shown as a function of Task (similarities, differences) and Context (low-bias, high-bias).

Number of Points Selected

We calculated the mean number of points that the participant selected across all of the 20 questioned signatures, separately for low-bias and high-bias trials. These data are presented in Figure 7.4. Because the two experimental groups performed different tasks, we did not directly compare this aspect of their data.

Similarities condition. A paired-samples *t*-test to compare the mean number of points of similarity selected in low-bias (M = 4.40, SD = 1.55) and high-bias (M = 3.92, SD = 1.64) trials revealed a significant difference, $t(91)$ = 4.56, p < .001, d = 0.30, 95% CI [0.27, 0.68]. Participants selected significantly more points of similarity in low-bias trials than in high-bias trials. There was no significant association between the mean number of similarities selected and response criterion in low-bias trials, Pearson's r = -.15, p = .14, or high-bias trials, r = -.01, p = .96.

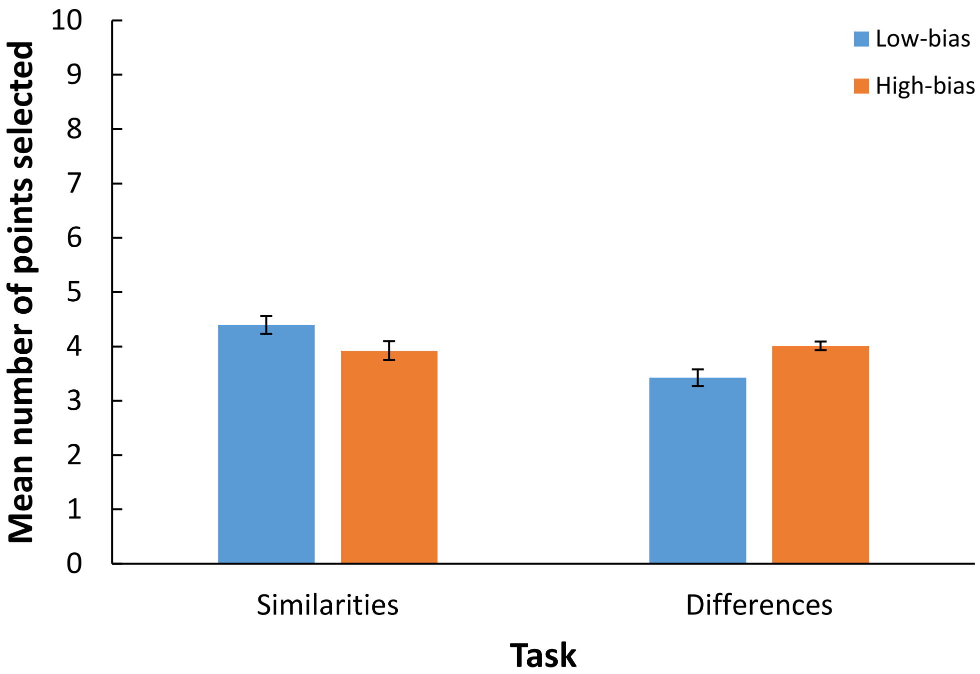


Figure 7.4. Mean number of points selected (±1*SE*) shown as a function of Task (similarities, differences) and Context (low-bias, high-bias).

Differences condition. A paired-samples *t*-test to compare the number of points of difference selected in low-bias (M = 3.42, SD = 1.49) and high-bias (M = 4.01, SD = 1.67) trials revealed a significant difference, $t(93) = 7.25$, $p < .001$, $d = 0.37$, 95% CI [0.43, 0.75]. Participants selected significantly more points of difference in high-bias trials than low-bias trials. There was a significant association between the mean number of points selected and response criterion in low-bias trials, $r = .32$, $p < .01$, and high-bias trials, $r = .26$, $p < .05$. That is, as the number of points of difference selected increased, so did response bias.

Discussion

There is mounting evidence for the presence of contextual bias in laypeople's judgements about the source of handwriting (Kukucka & Kassin, 2014; Miller, 1984; Study 1B). In Study 2, we investigated the extent to which contextual bias can be eliminated by altering the way in which our participants were instructed to complete the comparison task. Specifically, if the context effect in Study 1B is due to participants using a positive test strategy, then we should be able to eliminate the effect by instructing participants to focus on the similarities in the signatures. We, therefore, repeated the procedure from Study 1B; this time, however, we explicitly asked half of our participants to identify similarities between the signatures, and the other half to identify differences.

As in Study 1B, sensitivity was significantly lower in high-bias than in low-bias trials, whereas response criterion was significantly higher in high-bias trials than in low-bias trials. Contrary to our expectations, however, forcing participants to focus on similarities did not eliminate the effect of contextual information on sensitivity. Although sensitivity was higher overall in the similarities condition than in the differences condition, sensitivity still decreased when the signatures were accompanied by high-biasing information—regardless of task. Likewise, having participants focus on similarities did not eliminate the effect of contextual information on response bias. In fact, here we observed the opposite effect to that anticipated: participants in the similarities condition actually exhibited a significantly *higher* response bias than those in the differences condition.

One possible explanation for this effect is that participants' evaluation of the evidence was affected by the perceived ease or difficulty of the task they had been assigned. That is, in the similarities condition, participants may have found it difficult to identify similarities between the questioned and known signatures—especially in high-bias trials—due to directional cues that the signature was forged. Participants in the similarities condition selected fewer points of similarity in high-bias trials than low-bias trials, and we would expect that focusing on similarities would reduce response bias, as similarities are indicative of the signature being genuine. However, the dissonance between the expectation of a forgery and the difficulty of the task given to them could have created a debiasing backfire effect (Sanna, Schwarz, & Stocker, 2002; Schwarz & Vaughan, 2002). Research on subjective accessibility experiences suggests that people's beliefs are influenced by the ease with which they can generate reasons that support that belief (Jackson et al., 2017; Schwarz & Vaughan, 2002; Wänke, 2013). Schwarz and Vaughn (2002), for example, observed that participants who were asked to recall six examples of their own assertive behaviour judged themselves as more assertive than participants who were asked to list 12 examples. The authors suggested that, because thinking of six examples would be easier than thinking of 12, this ease could have been misattributed to the frequency with which they had been assertive (e.g., "If it's easy to think of examples, then I must be assertive"; Schwarz & Vaughan, 2002). In Study 2, the relative difficulty of the task—and the challenge of finding similarities in the face of information suggesting a forgery—could have led the participants to infer that there is little to no support for a 'genuine' conclusion and made them more convinced that the questioned signature was forged.

A follow-up study in which participants are asked to 'think aloud' during the examination process could shed some light on this possibility. Osborne, Taylor, and Zajac (2016b) used a think-aloud paradigm to explore the role of contextual information in bloodstain pattern analysis. Analysts in that study were given a photograph of bloodstains at a crime scene and asked to generate a hypothesis regarding the mechanism of stain deposition, while verbalising their thought processes. In particular, they were asked to talk through their internal reasoning for including or dismissing possible mechanisms. Analysts were also given the opportunity to request additional information, providing an updated working hypothesis after receiving each item. By asking their participants to think aloud, Osborne et al. (2016b) were able to better understand how contextual information directly influenced their participants' analysis. We suggest such an approach might also be beneficial in a follow-up to Study 2. If the challenge of finding similarities in the face of information suggesting that the signatures differed did indeed push participants towards thinking that the questioned signature was a forgery, we should see evidence of this in participants' narratives.

Regardless of the task we assigned to our participants, there is evidence to suggest that the contextual information affected the way in which our participants examined the signatures. Participants in the similarities condition identified more points of similarity in low-bias trials, whereas participants in the differences condition identified more points of difference in high-bias trials. These response patterns are likely due to expectations based on the directional cues in the contextual information (Risinger et al., 2002). Research suggests that, after making a prediction, people tend to search for information that supports—rather than conflicts with—their prediction (Windschitl, Scherer, Smith, & Rose, 2013). Participants in the differences condition may have found it easier to identify differences in high-bias trials than low-bias trials because they were able to search for features that supported their expectations. In contrast, participants in the similarities condition may have found it easier to identify similarities in low-bias trials than high-bias trials because, in low-bias trials, there were no context-driven expectations to interfere with their task. In other words, the contextual information may have created context-driven expectations in our participants that influenced the ease with which they could carry out their task (Saks et al., 2003).

There were several limitations in our current study. First, we only explored the potential for bias in one direction. One way to strengthen the study would have been to use bidirectional cues—that is, to have some contextual information that pointed towards a forgery, and other information that pointed towards the questioned signature being genuine. One way of doing this would be to capitalise on the fact that forensic examiners sometimes encounter contextual cues that suggest the suspect is not the source of the evidence found at the crime scene (e.g., that the suspect was in police custody at the time of the crime), and studies in other disciplines of the forensic sciences have found evidence to suggest that people can be biased both towards both guilt and innocence (e.g., Dror & Charlton, 2006; Stevenage & Bennett, 2017). We suggest a follow-up study to our current signature study which includes an additional context manipulation suggesting that the questioned signature is genuine.

Second, our design was between-subjects; one group was directed to look for similarities and another group was directed to look for differences. In reality, a document examiner should be performing both of these tasks (Koppenhaver, 2007; Morris, 2000). While Study 2 afforded us a rigorous way of considering the two tasks independently of each other, a follow-up study could ask participants to identify both similarities and differences in the same trial, assessing the number of each that are selected. Based on the data from Study 2, we would expect participants to identify more differences—and fewer similarities—when contextual information suggests a forgery (and vice versa when the information does not). Considering the signal detection data from Study 2, we would expect that allowing participants to focus on both similarities and differences would reduce the effects of contextual information on sensitivity and response bias, but not eliminate it entirely.

A final limitation is that although we controlled for the perceived complexity of signatures, we did not check that the 20 questioned signatures contained a similar number of similarities and differences. Because we used different signatures in the low-bias and high-bias condition (i.e., the signatures were not counterbalanced across the context manipulation), there is a chance that signatures in the high-bias condition had more differences than signatures in the low-bias condition. This would have made the task even more difficult for participants in the similarities condition and contributed to the debiasing backfire effect. A follow-up study could control for this by fully randomizing the signature and contextual information pairings. However, it is currently not possible to do this with the heat-mapping software that was required to run the experiment.

Where to from here?

The results of Study 2 suggested that contextual bias cannot be eliminated simply by altering the signature examination process so that participants adopt a negative test strategy. Although considerably more research is necessary to confirm this notion, in Study 3 we turned our attention to exploring ways in which we can manage forensic examiners' exposure to contextual information.

CHAPTER EIGHT

STUDY 3: WHAT ARE QUESTIONED DOCUMENT EXAMINERS' PERSPECTIVES ON MANAGING CONTEXTUAL BIAS?

Concerns of forensic examiners being biased by contextual information are by no means new. In fact, they were first addressed by Hagan (1894) in his treatise on handwriting identification:

> The examiner must depend wholly upon what is seen, leaving out of consideration all suggestions or hints from interested parties; and if possible it best subserves the conditions of fair examination that the expert should not know the interest which the party employing him to make the examination has in the result. Where the expert has no knowledge of the moral evidence or aspects of the case in which signatures are a matter of context, there is nothing to mislead him, or to influence the forming of an opinion. (p. 82)

More recent discussions on bias in forensic science led the National Research Council (NRC; 2009) to recommend that forensic laboratories develop robust standard operating procedures that minimize or mitigate potential bias and sources of human error by blinding the examiner to contextual information.

Not everyone in the forensic community, however, has been so willing to accept that bias is a significant problem in forensic science. For example, some practitioners still argue that forensic examiners are immune to bias or that bias can be overcome through sheer willpower (Butt, 2013; Leadbetter, 2007; Thornton, 2010):

> I reject the insinuation that we do not have the wit or the intellectual capacity to deal with bias, of whatever sort. If we are unable to acknowledge and compensate for bias, we have no business in our profession to begin with, and certainly no legitimate plea to the indulgence of the legal system. We can deal with bias, but we must work at it. (Thornton, 2010, p. 1663)

Such resistance is likely to exist for two reasons. First, an established body of literature demonstrates that although people can easily recognize and identify bias in others, they have far more difficulty recognizing when their own judgments and inferences are affected (e.g. the *bias blind spot*; Pronin, Lin, & Ross, 2002). As a result, people tend to believe that they are less susceptible to bias than those around them. When asked to complete a series of self-assessments, for example, people tend to rate themselves as less subject to various cognitive and motivational biases than others (Pronin et al., 2002), and actually tend to overestimate bias in others’ judgments (Kruger & Gilovich, 1999). Experts are not immune from this phenomenon (Steinman, Shlipak, & McPhee, 2001). In a recent survey on forensic examiners’ views on bias, 71% of respondents agreed that bias was a cause for concern in the forensic sciences, but only 26% of respondents believed that their own judgments were at risk of being biased (Kukucka, Kassin, Zapf, & Dror, 2017).

Unfortunately, being told about bias does not appear to be a protective factor. For example, in Pronin et al. (2002), participants showed a better-than-average bias, rating themselves as better than average on positive qualities and worse than average on negative qualities. Reading an explanation of the better-than-average bias did not reduce the effect; participants still insisted that their self-assessments were accurate and objective. Even if we admit that our own biases might pose a problem, they are extremely challenging to correct because most are out of our conscious control (Frantz, 2006). Wilson et al. (1996), for example, forewarned participants about the anchoring-and-adjustment bias in an attempt to eliminate it, but participants still showed the bias—even when motivated to be unbiased. That is, people appear unable to correct for their own biases (Frantz, 2006). Bias in forensic science, then, is unlikely to be overcome through willpower alone (Dror, Kassin, & Kukucka, 2013).

The second reason that discussion of contextual bias has been met with resistance from practitioners is that many forensic examiners *need* at least some contextual information to be able to perform the tasks required of them. Bloodstain pattern analysts, for example, rely on contextual information to put a crime scene in context and to determine what the purpose of the examination is (Osborne, et al., 2016b). There are concerns that withholding such information from the examiner could cause more problems than it solves (Thompson, 2011). Relying on information from an investigator who has no forensic science expertise, for example, could result in the examiner conducting one or more tests that are not appropriate for the propositions that should be considered (Thornton, 2010), and in a failure to conduct tests that are necessary (Thompson, 2011; Thornton, 2010).

For this reason, it is critical to understand—and distinguish between—the types of contextual information that are *task-relevant* and *task-irrelevant* to an examiner. The National Commission on Forensic Science (NCFS; 2015) defines task-relevant information as information that is necessary for drawing conclusions: about the propositions that the forensic examiner is asked to consider; about the physical evidence; and "through the correct application of an accepted analytic method by a competent analyst" (p. 2). In contrast, information is considered task-irrelevant if it is not necessary for drawing conclusions about the propositions the examiner is asked to consider, if it leads the examiner to draw conclusions from something other than the physical evidence, or if it "assists only in drawing conclusions by some means other than an appropriate analytical method" (NCFS, 2015, p. 2).

Whether a piece of information is task-relevant or task-irrelevant, then, is highly dependent on the examiner's task. According to the NCFS (2015), forensic examiners can be involved in three distinct phases of a criminal investigation. In the *preliminary* phase, the examiner surveys the crime scene, decides which evidence to collect, and determines the examinations that are needed. In the *analytic* phase, the evidence is examined, analysed, and compared. In the *evaluative* phase, the examiner helps the client understand the findings and puts those findings into context (NCFS, 2015). It is important to note that information can be task-relevant in one phase but task-irrelevant in another. For example, eyewitness statements might be important for determining where to search for fingerprint evidence (preliminary phase) but should not be considered when comparing latent and known fingerprints (analytic phase) (NCFS, 2015).

Within each phase, the different questions or propositions that forensic examiners are required to opine about will further inform decisions about the task-relevance of the information. For example, forensic examiners are often asked to consider different propositions about the source of the evidence (e.g., "do the latent and known fingerprints have a common source or different sources?"; Dror & Cole, 2010; NCFS, 2015). In this task, information about the surface on which the latent print was found would be task-relevant as it could help account for observed differences between the latent and known prints. Any other information, however—the suspect's criminal history or other laboratory results, for example—would be considered task-irrelevant (NCFS, 2015).

Even when the task and proposition are held constant, individual examiners' perceptions of what is task-relevant can vary. Osborne et al. (2016a), for example, asked bloodstain pattern analysts about the types of information that they believed they needed to conduct their analysis. Analysts were asked to classify a target bloodstain pattern (e.g., cast-off, drip stain). They conducted their initial classification in the absence of any contextual information, but were then allowed to request individual pieces of information that they considered necessary to refine their opinion. Available contextual information included a police briefing, medical findings, nearby bloodstain patterns, DNA findings, other forensic science evidence, and a witness statement. After receiving each piece of requested information, the analysts were given the opportunity to revise their opinion. This process continued until the analyst decided that no further contextual information was necessary, or until all six items had been received.

Although all bloodstain pattern analysts carried out the same task, there was considerable variation in the type of information that they considered task-relevant, and in how that information affected their opinions. For example, almost all of the analysts considered knowledge of other bloodstain patterns to be task-relevant, with 97.4% of analysts requesting the information, but this information led to a shift in opinion in only 57.8% of the analysts. In contrast, medical findings were requested by 92.3% of analysts and led to a shift in opinion in 90.0% of analysts. Overall, all analysts requested at least one item of contextual information, and 49% of the analysts (N = 19) requested all six items. Data analyses revealed that the final opinion given by 90.0% of the analysts had been influenced by contextual information. Although 70% of the analysts made correct pattern classifications, analysts who requested fewer items of contextual information were more likely to give a correct conclusion, whereas analysts who requested more items of contextual information were more likely to give an incorrect conclusion (Osborne et al., 2016a). Notably, there appeared to be a difference in what analysts considered task-relevant and what is task-relevant according to the NCFS (2015). In fact, according to the NCFS (2015), *none* of the available contextual information was relevant to the proposition the analysts had been asked to consider.

Contextual Information Management

How, then, do we grapple with the question of how to mitigate contextual bias, while still allowing practitioners access to the information that they need? Although there has been general agreement that there is not a “one size fits all” solution to this problem, forensic practitioners and academic researchers have proposed several *contextual information management* strategies—some of which have been implemented in forensic laboratories around the world (Budlowe et al., 2009; Dror et al., 2015; Found & Ganas, 2013; Krane et al., 2008; Osborne & Taylor, 2018; Saks et al., 2003; Stoel et al., 2014; Thompson, 2011; Wells, Wilford, & Smalarz, 2013). The most commonly discussed of these are described below.

The case manager model. In this model, the forensic examiner is assigned the role of either *case manager* or *examiner* (Found & Ganas, 2013; Stoel et al., 2014; Thompson, 2011). The case manager communicates with investigators (or attorneys, in civil cases), provides instructions on which evidence needs to be collected at the crime scene or from the suspect, and manages the overall workflow of the laboratory. The case manager determines which information is task-relevant, and passes only that information on to the examiner. This separation of function allows case managers to remain aware of the investigative context, while the examiner remains blind to task-irrelevant contextual information and is thereby safeguarded against contextual bias. The case manager approach, therefore, appears to provide an appropriate balance between the examiner knowing too much information and not knowing enough (Thompson, 2011).

The Netherlands Forensic Institute (NFI) recently implemented the case manager model in its firearms department (Mattijssen, Kerkhoff, Berger, Dror, & Stoel, 2016; Stoel et al., 2014). All incoming cases are now scanned for relevant information by a first examiner; a second examiner then works on the case having received only the task-relevant information from the first examiner. Mattijssen and colleagues (2016) reported that the context management scheme has shown very promising results, and the NFI is currently trialling the same scheme in questioned document examination.

In 2009, Found and Ganas (2013) introduced a context-manager model in questioned document examination at Victoria Police Forensic Services Department. They determined that several pieces of information were invariably task-irrelevant: the "investigator's rank, the squad where the investigator works, the offence type, the description of the incident, the summary of charges, and any description of any admissions made in relation to the case" (p. 156). Cases are stripped of all such information, leaving only what is necessary for carrying out examinations and comparisons—in this case, the known and question populations and the dates associated with the known and questioned writing. The scheme was implemented in

2009, and at the time of publication in 2013 the department had not noted any negative outcomes.

Sequential unmasking. Sequential unmasking optimizes the order in which an analyst examines the questioned and known samples. Side-by-side comparisons of the questioned and known samples leave examiners vulnerable to circular reasoning, in which an examiner's interpretation of the questioned sample might be influenced by features found in the known sample (Dror et al., 2015). As a result, an examiner might identify important features in the sample collected from the suspect and try to match these to features in the questioned sample, increasing the risk of confirmation bias (Dror, 2013). In sequential unmasking, evidentiary samples (e.g., questioned handwriting) are examined first and reference samples (e.g., known handwriting) second. In this way, the known material cannot "suggest" features for the examiner to find in the questioned material (Krane et al., 2008). The DNA department of the NFI is currently using the sequential unmasking method in their laboratory; again, results have been encouraging (Stoel et al., 2014).

Blind/independent review. Forensic laboratories often require conclusions to be peer-reviewed by a second examiner before a case report can be finalised (Ballantyne, Edmond, & Found, 2017). Several researchers have suggested taking this procedure a step further, by making this procedure a *blind review*, in which the person reviewing the findings does not have access to potentially biasing case information or the initial examiner's findings (Budlowe et al., 2009; Dror et al., 2015; Osborne & Taylor, 2018). This way, even if the initial examiner had access to the case information, the second examiner's examination will remain unaffected by possible contextual bias. Additionally, blinding the reviewer to the first examiner's conclusions will protect them from any expectation-driven biases (Osborne & Taylor, 2018). Supporters of this strategy suggest that it might be more practical to implement blinding during the review process rather than the initial examination, but whether this is true has yet to be thoroughly investigated (Budlowe et al., 2009).

Evidence line-up. Saks et al. (2003) argue that current evidence-matching practices are the equivalent of an eyewitness 'showup,' in which a witness is shown one photo and asked if it is the perpetrator. A considerable body of literature shows that this "one option only" process significantly increases the chances of mistaken identification when compared to a line-up, in which photos of known innocents are presented with the suspect's photo (Steblay, Dysart, Fulero, & Lindsay, 2003). To overcome this problem, Wells et al. (2013) have proposed that forensic examiners use the *filler-control method*—essentially a forensic evidence line-up. In this method, the forensic examiner is first presented with the evidence sample from the crime scene, followed sequentially by at least two comparison samples—one from the suspect and the other from a known innocent (or *filler*). There is evidence to suggest that an evidence line-up could protect against contextual bias. In Miller (1984), participants who evaluated the evidence in this way appeared unaffected by the contextual information.

The filler-control method could be particularly beneficial in forensic disciplines where the examiner needs to know the contextual information for the purposes of the examination, or the information is embedded in the evidence itself. That is, even if the forensic analyst has been exposed to contextual information, they will not know which of the two comparison samples came from the suspect, and therefore cannot rely on the contextual information to inform their opinion.

The procedures described above show how some forensic science agencies could take—or have taken—steps to reduce contextual bias in their examinations. Furthermore, when such procedures have been implemented, they have yielded promising results (e.g., Found & Ganas, 2013; Osborne & Taylor, 2017; Stoel et al., 2014). Relevant to this thesis, however, are questioned document laboratories specifically.

Unfortunately there are no data available on the number of questioned document laboratories using contextual information management strategies; neither do we have comprehensive data on which strategies are being used. Furthermore, discussions about contextual bias in questioned document investigation tend to overlook private and sole-practitioners. We also note that some of the recommended contextual information management strategies proposed in other forensic science disciplines may not be feasible for document examination. We considered these issues in Study 3.

The Present Study

This study was part of a project funded by the National Institute of Justice (NIJ) in the United States. The purpose of Study 3 was to understand the practitioner's view of contextual bias, current contextual information management procedures that are being used in practice, and any barriers that practitioners face in implementing strategies for managing contextual information. We interviewed questioned document examiners using a semi-structured question schedule, and analysed participants' responses using a deductive thematic analysis approach (Braun & Clarke, 2006).

Method

Participants

The participants were 19 professional document examiners from both government laboratories (n = 11) and the private sector (n = 8). Participants' years of experience ranged from 7 to 48 (M = 23.37, SD = 12.45). All participants were current practitioners and certified in accordance with standards from the American Society of Questioned Document Examiners (n = 3), the Australasian Society of Forensic Document Examiners (n = 7), the American Board of Forensic Document Examiners (n = 5), the National Association of Document Examiners (n = 2), the Southwestern Association of Forensic Document Examiners (n = 1), or the European Network of Forensic Handwriting Experts (n = 1). All participants were informed of the purpose of the study and how the data would be used and stored; all gave their written consent to participate.

Procedure

Two experimenters conducted the interviews either in person, via Skype, or via telephone. Ten of the interviews were conducted by both interviewers; the remaining nine were conducted by one of interviewers due to interviewer availability. The interviewers followed a semi-structured interview protocol (see Appendix B) according to six specific research questions:

1. What are document examiners' general views on contextual bias?
2. What types of contextual information can document examiners be exposed to and what are the sources of this information?
3. Which types of contextual information are task-relevant and task-irrelevant?
4. Do document examiners ever experience pressure to reach a specific conclusion? If so, from what sources?
5. What contextual information management strategies are document examiners currently using?
6. What barriers do document examiners face when considering contextual information management?

Prior to the interviews, all participants were sent an information sheet which detailed the purpose of the study, the aim of the project, what was expected from the participants, and the data that would be collected. Participants were also sent a consent form to sign and return before the interview took place.

Analysis. The interviews were audio-recorded, transcribed verbatim by a professional transcription service, checked by the experimenters to ensure accuracy, and then coded according to a deductive thematic analysis protocol (Braun & Clarke, 2006). We organized the data according to themes at the surface level (semantic content), and used a deductive, semantic approach to assign codes to segments of text relating to the six research questions.

The analysis comprised three main stages. First, the transcripts were read through several times and broad subjects and phrases of interest were identified. Second, the initial notes were refined into more specific topics that related directly to our research questions. Third, relevant data extracts were transferred into an Excel worksheet and grouped and coded according to the themes identified. Two experimenters then reviewed the coded data extracts together and evaluated the appropriateness of the assigned codes.

Results

A summary of examiners' responses to our research questions is outlined and discussed below, accompanied by illustrative quotes[10].

General views on contextual bias

All of our examiners reported being aware that biased evaluations of the evidence due to contextual information is an issue that is being discussed in the forensic science community. The majority recognized that forensic examiners should be aware of the risk of bias, but views on the extent to which contextual bias is problematic varied. Some examiners saw bias as something that could not be controlled—that just because they try to be unbiased, it does not mean that they are. One document examiner noted that the biases they are unaware of pose the greatest risk to their objectivity. Some examiners stated that they saw bias in themselves and their co-workers, and would often question their own objectivity and try to minimize the risk for bias as much as possible. Some examiners expressed an explicit motivation to not be biased—arguing that if they allow the contextual information to inform their conclusion rather than the evidence, it could reflect poorly on them in court and cost them in the future.

> I think that it has certainly been demonstrated that it [bias] can be an issue, and it's something that perhaps people may not have been aware of in the past, but having an awareness of potential influences is something that has been brought to light.
> (QDE 16)

[10] All quotes have been de-identified and minor grammatical corrections have been made to improve readability. Note that QDE refers to questioned document examiner.

> You're naïve if you don't think bias is playing a part, of course, because you're a human being, but you do try to feel conscious of it and to mitigate against it. (QDE 19)

> The concept of unconscious bias comes into play. You have to try not to anchor to what you have been told… We have to be careful, because just because we try to divorce ourselves from bias doesn't mean we do. (QDE 10)

> I guess the problem with these kinds of biases is that it's the stuff that you're not aware of that's going on internally that's the problem. (QDE 17)

> Do I have bias? Absolutely. And do I have to keep my focus away from that and sometimes think while I'm working on a case, "am I really looking at this from an objective perspective?" "Am I really being objective?" Because when someone tells you something, you can't help but anchor to that, unfortunately, and you have to really try to divorce yourself and be able to do an introspection on that… Do I see bias in other examiners? Absolutely. (QDE 10)

> If I can't go there and support [my opinion] with demonstrable testimony there's no point in me saying anything because another document examiner will come in and shoot everything I've said down. I'm not doing anybody any favours. I'm wasting people's time and I'm going to look like a fool. Why would I want to do that? (QDE 4)

In contrast, other examiners attributed erroneous conclusions to practitioner incompetence rather than bias, and believed that their expertise should be sufficient to protect against bias. These examiners also argued that contextual information would not bias them because they did not take it into account when coming to a decision.

> It irritates me, to be painted with the same brush of the people who take shortcuts or weren't trained properly… There are so many features that support a finding, that individuals who are not doing a thorough job can be influenced by anything. (QDE 8)

> I have personal experiences that helped me discount any time any information that points to an identification or elimination or not. The evidence is going to show or not that the writer did or did not write something. The issue of bias, to me, is misplaced. (QDE 8)

> To me it doesn't matter where it comes from. If [the suspect is] in jail, they're in jail. I'm just going to look at it as objectively as possible because when you're looking for similarities or dissimilarities, I have to physically find those things and mark those

> areas. I can’t just mark whatever out of thin air. In that sense, it doesn’t matter what they say in the notes. (QDE 6)

What types of contextual information can document examiners be exposed to and what are the sources of this information?

Document examiners reported encountering contextual information from three main sources: (1) their client (attorney or investigator); (2) their work environment; and (3) the evidence itself. The different types of contextual information are considered in turn below.

Case information. Document examiners reported that they were regularly given case information by their client—including what is being alleged, background information on the individuals involved in the case, and the history of the questioned document. Examiners reported receiving this information either via a letter of transmittal, via email, or through a phone call. Examiners who worked in government laboratories reported having access to case information through administrative channels, such as a Laboratory Information Management System (LIMS) or evidence submission forms.

> They want to give me the whole story. I mean it’s just on and on and on, who in their family hates who, and their mother was being taken care of by their sister who was stealing money from the bank vault. They want to give me all of that. (QDE 15)

> Often, about 50% of the time, I’ll get text within that cover letter that says, “so and so was witnessed by a notary and two cousins watching him sign this” (QDE 8)

> We have this LIMS system… Right now I can go in and look at anybody’s opinion. Even with the regular relevant information or irrelevant, whether this person is a victim or is a suspect, that’s in there. In DNA they put a whole story of what happened… That could be in there. (QDE 3)

> It is actually on the form quite explicitly. It says: Indicate whose premises the nature of the contact between the suspect, volunteer, victim, witness, scene, whether there was prior opportunity for contact. The owner, source of the items, nature of injuries to victim, suspect, and who bled. List any other known aliases in this section in relation to drug cases of the audience below located at a previously attended [clandestine] scene, which are of particular significance to the investigation. It's actually taking quite a lot of detail. (QDE 16)

Theory. Document examiners reported experiences in which they were told their client's theory regarding the questioned document, as well as what the expected outcome of the examination was. Examiners who worked in government laboratories reported experiences in which investigators would try to tell them how the investigation is going and would share their own theories on the case.

> The person whose will it was supposed to have been had died and his two tenants in his apartment and their attorney were suspected of having forged the will and taken charge of his assets. (QDE 1)
>
> Just a lot of persona context that we just don't need to know, so like, "he was in the area at the time, and so it's gotta be him" (QDE 12)
>
> Sometimes we get the theory. The whole, you know, "my guy forged this and I just need you to tell me" (QDE 5)

Physical condition of the writer. Document examiners reported instances in which they were informed of the physical state the writer was in at the time the writing was allegedly created—including the writer's age, whether they had any physical or mental impairments, or whether they were intoxicated.

> If there's tremor, usually you can tell in that person's writing, but sometimes the evidence of age overlaps the evidence of intoxication or other types of affliction… (QDE 3)

Information about the questioned document. Document examiners also considered the nature of the document itself, for example a will or a threat letter, to be contextual information. According to our document examiners, the type of evidence to be examined can indicate that the suspect's writing sample was produced in prison.

> That is always the case with letters from jail, and threats… There's potential to have bias in judging that. (QDE 12)

> I've also had situations where - US jails here, there's something called Tank Orders. Tank Orders are a form that a prisoner will use to order shaving supplies or they'll want money taken out of their inmate account to buy new socks or something like that. These tank orders are saved and have been used as known writings many times. Obviously, when you usually do this often enough, you know what a tank order form looks like, so you know the handwriting is coming from somebody who's either been incarcerated or is even currently incarcerated. (QDE 2)

Other evidence. Examiners reported instances in which investigators told them that a suspect had confessed, that there was other evidence that incriminated the suspect, or that circumstantial evidence suggested the suspect wrote the questioned document.

> Saying, "well this person confessed", I had that repeatedly, repeatedly during my years as a government examiner. (QDE 8)

> Police officers saying they've got CCTV footage of the person signing the check in the bank. (QDE 13)

> A couple of occasions we're left with two possibilities. One of whom was the captain of the boat, and the other of whom was the cook and it was put to me to say this was found on the bridge, who is it likely to be? I said, that's not for me to determine, they're then able to say, the captain has access to the bridge, the cook didn't. (QDE 13)

Criminal history. Some of the examiners reported experiences in which they were told that the suspect had a criminal history of forgery. Some also reported that the comparison material could indicate that the suspect had a criminal history.

> That's a very classic form of context biasing, "this guy did 10 years behind bars for forgery, and now we've got this other case, and would he do it again?" (QDE 2)

> One of the things that would happen is that we would use as known signatures of the person on the fingerprint cards from prior arrests. We would have the ... Use the information on the fingerprint card, if it was a kind of situation where they filled out their own height, weight, that sort of thing. We would use their imprinting, but at the least, at the minimum, they would have signatures. (QDE 2)

Which types of contextual information are task-relevant and task-irrelevant?

Some of our examiners stated that contextual information was not relevant in the examination of questioned documents; they believed that all they needed to know was the purpose of the examination and what the propositions were. Others, however, reported that some types of contextual information could be relevant, and that these differed on a case-by-case basis.

> I think from our perspective, document examination seems to lend itself more to not having any of that information available, because generally we just need to know what needs to be examined, so what is in question, and what is being supplied for comparison purposes. We're in I think a good situation where we cannot rely on any of that additional information. (QDE 16)

> If we do accept the case, then on a case by case basis we have to know, I think, what's being alleged… Either the signature is being questioned, so the potential that the document has been altered has to be considered… So, I need the information about the history of the document and what led up to the situation where the document is being questioned in some way. We might need to have other information about the physical condition of the writer. (QDE 2)

Task-relevant information. Many of our examiners believed that it was important to know the history of the document that they were examining. Knowing when the document had been created and what writing instruments had been used to create it was considered helpful to establishing whether the questioned and known writing are contemporaneous, which is important for deciding whether it is appropriate to conduct a comparison. According to our examiners, if the questioned writing was produced 20 years ago and the known was written one year ago, then it would be inappropriate to conduct a comparison due to likely changes over time. A knowledge of the tools used to create the writing was also reported as potentially task-relevant; knowing the type of pen used, for example, could be relevant if the pen used to write a signature did not exist at the time the document was allegedly created.

Knowledge of factors that could have affected the physical appearance of the writing was also reported to be task-relevant. According to our document examiners, the surface on which a document is written can affect the appearance of the writing—writing on a wall, for example, will look different from writing on paper. Document examiners also reported needing to know information about the writer that could explain distorted or unnatural writing. A person's writing, for example, could vary depending on the position in which the writing was done (e.g., sitting at a table, on a train, or propped up in bed on a pillow) or a writer's medical condition. Many document examiners noted that being ignorant of these facts could affect their ability to assign importance to differences and similarities in the writing.

Lastly, it was reported that the nationality of the writer could be task-relevant, as it could help to establish a reference population, which is critical when considering the rarity of features in the writing.

> It's important for document examiners to be somewhat familiar with that so that they don't erroneously interpret the strange artefact on a document as something that's the result of forgery or an attempt at creating non-genuine document versus side-effects from PDF technology. (QDE 8)

> When I was in training, we had a case come in and the writing just looked really odd. It was, it looked like somebody had just tremors, but it looked not like natural tremors, but it didn't look like tremors of forgery. It just was a strange, tremulous writing… It turned out that the guy had written it on a patio table, a concrete patio table. (QDE 11)

> Sometimes there are cases of a specific nature, so let's say there's an anonymous note and it's written on a wall. I'd like to know how high on the wall was it written or maybe something that would give me some ideas about, if there's any report about the substrate or the circumstances of the signing, then it would be useful to know that at some point. (QDE 15)

> Do you happen to know the writing conditions when this signature was reportedly signed? For example, was she propped up in bed on a pillow? Was she at a table? Was she sitting in a wheelchair? That all has a very important impact on the way your signature can be unintentionally distorted. (QDE 7)

> We need to know whether or not the person has a degenerative disease, like Parkinson's or something that would affect the handwriting over a long period of time, in between the time that the document was allegedly signed and today when we're given the exemplars. (QDE 2)

> If they have information that the person has some type of affectation that affects their writing, we want to know that upfront because that is going to affect the exam. (QDE 5)

> My initial opinion was "yes your friend wrote that", that was my initial opinion, but then I contacted a document examiner in Poland and said can you send me a copybook from Poland … It turned out what I was seeing as consistent anomalies between the questioned and the known, was the Polish copybook. So that would be expected, so my opinion completely changed. Once I took that into consideration, that's contextual information, but I wanted it. Because even though they're both Latin-based alphabets, they have a different copybook than we do. (QDE 11)

Task-irrelevant information. Information is considered task-irrelevant if it could lead an examiner to draw conclusions based on something other than the data contained in the writing material (NCFS, 2015). All examiners agreed that any information not related to the physical aspects of the document is task-irrelevant.

> I think more often than not you can get away with Q and K and ignore everything else. (QDE 13)

When asked about specific types of contextual information that would be task-irrelevant, document examiners noted the following: the case background, their client's theory about the source of the questioned writing, the criminal history of the suspect, and other forensic science evidence.

> The whole history of the family, who's related to whom, why Sally cannot stand Jim, and Christmas dinner was the worst… [They say] "That's why we know that she forged it, because they just had a big fight at the dinner table", and oh my god, [they] just have to go on and on. (QDE 11)

> That [theory about the source] is one of those things that I really don't want to know and will try to stop the attorney, but I'm usually aware just logically that if one side feels the document is forged, the other side is proffering the document because they believe it's genuine. (QDE 2)

> [The criminal history] is something that I wouldn't necessarily need to know or want to be told because that's … to me that's highly biasing. (QDE 1)

Do handwriting examiners ever experience pressure to reach a specific conclusion? If so, from what sources?

Some examiners described experiences in which they felt pressured by their client or workplace to reach a specific conclusion. Examiners who worked in the private sector described instances in which the attorney misunderstood the nature of document examination, and tried to convince the document examiner that, for example, one point of difference was enough to exonerate their client. Document examiners who worked in government laboratories reported feeling considerable pressure at times, or even being confronted by the investigator if they did not reach the desired conclusion. These experiences would sometimes result in an investigator trying to change the document examiner's opinion—either by questioning the examiner's competence or going through the evidence with the examiner and pointing out all the features that supported the desired opinion. Some document examiners reported having been lied to by the investigator, in an attempt to influence their opinion.

> They might come back and sort of suggest, "have you seen this?" I've been called a liar. I've been called incompetent because the result didn't go the way that that person wanted. That's got nothing to do with the examination or my competence. That's just got to do with their interpretation of what they wanted the result to be. (QDE 14)

> [The investigator] said, "We just need you to say…", the old "we just need you to say" (QDE 4)

> Years ago [I] had a case where the investigators gave me something and said, "This writer has confessed to writing these cheques". I examined and I said, "You cannot get to this writer on these cheques. I don't think they wrote them". They were angry with me because they had a confession and I was screwing up their case. (QDE 4)

> I've had them argue. I've got a case out there where a District Attorney wanted it to be his defendant and not the victim's writing. I started showing him, "Well, it's this, this, and this. This is why it…" He goes, "what about that? It's different". There was a lot of argument over that. (QDE 4)

> I’ve had a detective say to me when I gave him an inconclusive, “well is there somebody else who is more of an expert than you that could…” [the inference] that maybe you’re just not as much of an expert. (QDE 3)

> Investigators will say, “What about…”. [My colleague] had this case just within this last week. [The investigator] “Look at the E. Look at that middle initial E.” Their threshold for identification is much lower than ours. (QDE 4)

> I’ve just personally observed too many instances where people confessed to something or investigators would lie when they submitted a laboratory examination request. They would say that someone admitted writing something, or that they watched another person writing something. (QDE 8)

What contextual information management strategies are handwriting examiners currently using?

Four contextual information management strategies have become popular with academic commentators: the case manager model (Found & Ganas, 2013; Thompson, 2011), sequential unmasking (Krane et al., 2008), blind/independent review (Budlowe et al., 2009; Osborne & Taylor, 2018), and the evidence line-up (Saks et al., 2003; Wells et al., 2013). When we asked our examiners whether they were currently using any of these strategies, we received mixed responses. It also became clear that document examiners who worked in the private sector had a different approach to contextual information management from those who worked in government laboratories.

The case manager model. Several examiners reported using techniques aligned with the case manager model, in which a supervisor or another document examiner would remove contextual information prior to the examination. Some of the laboratories our examiners worked in would only screen out the case circumstances included in the evidence submission form, whereas other laboratories would remove any and all extraneous case information. One reported issue with this approach, however, was that cases took longer to complete due to the extra work involved in removing the information.

> We have tried to implement a strategy where someone other than the person who is going to examine the case will go and do the vetting, or they will screen out the information that is on the [case summary] page. (QDE 16)

> Both directors vet all our cases and are responsible for removing contextual information that's not necessary to the case before it comes to us. They will say what time have you got, discuss resources, discuss what they might need to collect in order to conduct the examination, or to give us the case. (QDE 12)

> The main thing that we have done is try to on our police PA187 submission form, is to have an examiner screen out. There's quite often a story that's put in about the circumstances of a case that exhibits come in for examination, which we screen out, so that is one thing that we've implemented. We have one examiner screen that information out, fill it into an envelope, so that there's a copy of that page at the back. If there's any pertinent information that is within that description, we would pull out that if it's not explained within what the examination requested is. (QDE 16)

Sequential unmasking. When asked about standard operating procedures for the examination process, our examiners' responses were mixed. The majority of examiners reported a process similar to linear sequential unmasking, where they would analyse the questioned writing first, then the known sample, before comparing the two. Other examiners stated that the order in which they examined the writing depended on the case—sometimes they would analyse the known sample first, and other times they would analyse the questioned sample first. One examiner noted that "the protocol"[11] does not state whether an examiner should analyse the known or the questioned first, and believed that the order is determined by the examiner's experience and what works best for them.

> Basically, what it says is that one examines the questioned writing first and it's… you're examining it not in detail but you're looking basically for naturalness. (QDE 1)

> Usually, in a handwriting case, I'll do a preliminary examination of the questioned document and start my notes, writing in those what we call in the laboratory trademark characteristics (QDE 2)

[11] The examiner did not identify a specific protocol, but since they were trained according to SWGDOC standards we assume that they are referring to the SWGDOC Standard for Examination of Handwritten Items.

> During an examination of writing, I will first, depends on the case, but quite often first examine ... Let's say it's a signature case. I will examine the known signatures. In other words, the non-disputed signatures that are submitted to me. The way I examine the known signatures I say, let's say there's 20 of them, I do a cross comparison of the 20 known signatures to make sure they were all written by one individual. (QDE 7)

> I tend to look typically at the question first to find out what are, are there really anomalies in there. Then I'll go look at the known, and then do a compare and contrast. (QDE 10)

> I'm not sure that the protocol really spells out whether or not you should do the known or the question first and I think it's a matter of your experience and what works best for you. (QDE 11)

Blind review. Some examiners who worked in the private sector reported having colleagues who would peer review their opinions; several of the sole-practitioners described having an agreement with other document examiners who review their work on a *pro bono* basis. Throughout our interviews, however, it became clear that it is not standard practice for private document examiners to have their work peer-reviewed. Document examiners who worked in a government or police laboratory reported having at least some of their work peer-reviewed, however the peer-review process was not standardized across laboratories. One examiner’s laboratory utilized blinded reassessment, in which the whole case was worked by a second examiner who was blinded to the findings from the first; other examiners’ laboratories performed technical or administrative reviews of examination notes and conclusions. Although peer-review appeared to be standard practice in government and police laboratories, it should be noted that the majority of these laboratories did not *require* the reviews to be blinded.

> When I have a case that is either confusing to me, or I think there’s something dicey about it, or maybe I even suspect my own ability to be objective about it, I have a colleague or two that I will send an image to and have a conversation with, but it’s nothing formal and it’s nothing that I report that I’ve done. (QDE 15)

> Any of the cases that I have reviewed, I have an agreement with several people, that "hey, would you take the time to review a case for me?" We generally will do it first just through email. Have good images, high quality images sent. (QDE 7)

> Our type of review is essentially verification, that's the verification process that you fill out a form and you're just verifying the handwriting, or does the pattern match the opinion. (QDE 3)

> Everything here is reviewed by another examiner. They will look at the notes. Typically what I do is I see what's in question and I will… Very much like an examination, I'll see, "Okay, what are we dealing with and what did this other examiner have as exemplars". I try to prove them wrong. (QDE 4).

> They wouldn't get to see the first examiner's notes… I don't see anything, including the detailed notes, and I certainly don't see the conclusion. (QDE 13)

Evidence line-up. None of our examiners reported using an evidence line-up. One examiner noted that this type of contextual information management is not feasible for document examination, as it can be very challenging to obtain appropriate comparison (known) material from just a single person. According to this examiner, creating fillers that would be convincing to a document examiner would take time and resources that laboratories simply do not have.

> We have enough trouble getting specimens from one person, let alone asking an investigator to create four other people who they don't think did it. We'd have to have an evidence management unit whose job it was to then go and try and source fake specimens as it were. If they're gonna do that, then they need to make it real, so it has to be fake specimens which kind of look like the writing that's in question, not stuff that's completely different. I don't think that's a particular runner. (QDE 17)

Additional contextual information management strategies. Our examiners also reported using strategies other than those proposed in the literature. Document examiners who worked in the private sector, for example, tended to employ strategies that focused on the risk of exposure to task-irrelevant contextual information (e.g., that it could be used against them in court by an opposing attorney), and therefore emphasised giving direct instructions to the client as to the information that was—and was not—task-relevant. Several private examiners described having a professional website, on which their clients could learn more about the examiner's services and contact them directly; these websites could contain an explicit warning to the client not to give the examiner too much information.

Other examiners who were usually contacted by phone reported giving the attorney a verbal disclaimer at the start of the call. Examiners reported trying to keep these phone calls short and transferring the conversation to email instead, which had the added benefit of ensuring that there was a written record of all the information provided. One of the examiners had considered having a disclaimer on their answering machine, but was yet to implement this.

> I have tried to mitigate this somewhat. If you go to my website, I have a paragraph on the home page that says roughly, "Before you call a document examiner, take a moment to think of how you could ask them a question without revealing the answer you hope to get". (QDE 15)

> [Disclaimer via phone call] "Counsellor, I've got to tell you something, that I've been through enough depositions where I can tell you for certain, with a certainty, that the very first question that I'm going to be asked after what's my name and where do I work, is, "What were you told about this case?" And if you start telling me what a jerk this other person is, I'm going to have to testify to that. So it's not going to be helpful at all to your client to bring this other information in"… and that works. (QDE 2)

If none of these strategies worked, the document examiners reported trying to avoid reading the case material whenever possible. They acknowledged, however, that this was not always possible to do—especially if the contextual information was contained within the questioned document itself.

> You can avoid reading it. You have to manage whether you would be influenced by it. You have to be conscious that you could be influenced by it. But you can't, in the nature of the work that we do, the physical written words are the things that we're looking at, so therefore the nature of that is that it's unavoidable that you will be aware of what the content of the letter is. (QDE 14)

Document examiners who worked in government laboratories reported providing specific instructions to investigators. Several examiners reported having to educate investigators on the risks of revealing too much information, saying that they would often interrupt an investigator who said too much, or ask the investigator to cancel the request and resubmit it again at a later date. Other examiners reported trying to limit interactions with investigators—and indeed with other forensic units—to reduce the chance of being exposed to contextual information. Working in a government laboratory meant that the document examiners sometimes shared a building or a floor with examiners from other forensic disciplines who might be working on the same case.

> We actually have asked police officers in the past who have given us way too much information, or over-shared in their narrative to cancel the job and resubmit it after some training. (QDE 12)
>
> It's quite often when they ring up and they want to speak to us, we'll get a full story and we'll cut them off and say, "We don't need to know these details". (QDE 16)
>
> I think that from our perspective, we try to limit the conversation that we have so we don't know too many details. (QDE 16)
>
> I wouldn't want [other forensic evidence] to influence my opinion, so I purposely don't. I make sure I don't go and interact with the other disciplines in the building, for example. (QDE 12)

What barriers do handwriting examiners face when considering contextual information management?

Document examiners who reported having implemented contextual information management were asked whether they had faced any barriers in doing so. We also asked those who had not yet implemented such strategies whether they planned to, and their reason for their response. The examiners gave consistent responses; those who had already implemented context management said they had found ways to overcome the barriers, and those who had not were either trying to work around barriers or were somehow prevented from implementing the strategies. It also became clear that private document examiners and government examiners faced different challenges in terms of managing contextual information.

Private sector. For those who worked in the private sector, and especially those who were sole-practitioners, the largest reported barrier was a lack of resources—time, money, and additional practitioners. These examiners noted that some strategies, such as blind review, require additional personnel that they simply do not have readily available.

> A lot of it I don't see how I'm going to implement, because I don't have a case manager and don't… I mean, people on the Human Factors Committee are talking about these schemes that could come into play, where there was some central agency that took in all the private cases and then distributed them or some scheme like that, but I don't see it as something that's all that practical or will happen any time soon. (QDE 15)

Private practitioners also reported barriers relating to confidentiality, such as being subject to nondisclosure agreements and, therefore, not being able to send questioned and comparison documents to another person to be reviewed. Even if they were able to have someone else working on the case, they reported that the second examiner would need to be paid. Some sole-practitioners reported overcoming this barrier by having an agreement with other private practitioners, in which they "traded" cases free of charge.

> Having more peer review is probably a good idea, but it's a problem when you do what I do, because you get these documents and first of all, it's all implicitly or explicitly confidential. You're not supposed to take these and just throw them around. Number two, any time you ask another expert to do an examination, they've got to be compensated, if they're really doing a full examination. (QDE 15)

Government sector. Compared to private practitioners, document examiners who worked in government laboratories reported a larger number of barriers to implementing contextual information management strategies. They reported being constrained in their ability to implement context management strategies due to resistance from supervisors and managers, and the feeling that their colleagues were unwilling to acknowledge their own propensity for error. As mentioned earlier, some even described experiences in which they were explicitly pressured by their client or workplace to reach a specific conclusion based on task-irrelevant information.

Another reported barrier was that the information is always readily available in the Laboratory Information Management System (LIMS). Ideally, examiners stated that they would avoid using LIMS, but they stated that this is not always possible as the system contains information that the examiner needs for administrative purposes (e.g., the case number). When asked whether the system could be changed to address this problem, one examiner mentioned that it would require considerable restructuring, and was not likely to happen due to a lack of endorsement from management.

Government document examiners also reported being subject to institutional policies that can be slow to change, and that their laboratories often lacked the resources—such as time, money, and people—to implement effective context management strategies. The examiners noted that implementing contextual information management strategies can put a strain on already limited resources, and there is uncertainty whether these strategies are successful in reducing bias. For this reason many of our examiners' laboratories have been hesitant to implement them.

> Well I know, when I first brought it up [our supervisor] had quite a few issues with it… I just heard an explosion of why it's bad, why it's not going to work. (QDE 3)

> People are very much, "No I don't need [context management]. I have a zero percent error rate. I've never made a mistake"… You hear about a mistake being made, "well it's just kind of par for the course, that we are human". I know, but shouldn't you investigate a little bit more, able to see and maybe there's something else to it. (QDE 3)

> There just doesn't seem to be an openness to change or to embrace any of this stuff … I mean whatever we do—quality assurance—has to be approved through … it's just, when there is no enthusiasm above, you know. (QDE 3)

> It's just you've got such a small pool of examiners it makes it difficult to have that where someone, if you take it in, you can know information that the examiner doesn't necessarily know or that you provide them only what they would need to know as a document examiner to complete the case. (QDE 4)

> It's nice to have research that says, "this is what you should do, and you should get someone else with the appropriate training to take out all the redaction, this and that, and take out all the context", but they can't, unless they're [document examiners], and yeah, [document examiners] don't grow on trees. (QDE 12)

> We've got no measure of how effective this is down the track, like whether it has made any difference or not. (QDE 16)

> I don't know how you measure whether that's having an impact … There sort of seems to be an agreement that it's a good thing, but I don't know how big the problem of contextual bias is. Until and unless we can build up a picture of bias, other than thinking it must be there and it's a really bad thing, it's hard to see, for me to see that it's a really great thing. I think it would create some problems, I mentioned resourcing. And I also think it's kind of … even the person stripping out the contextual information is going to have some biases. (QDE 19)

Regardless of agency type, examiners noted that contextual information was sometimes embedded in the evidence itself. For example, the questioned writing could contain incriminating information, threats or vulgar language, information that could elicit an emotional response (e.g., bloodstains, a suicide note). Our examiners argued that it would be difficult—or even detrimental—to the examination to remove this information because it could also contain important unique characteristics of writing. In this way, examiners perceived a distinct trade-off between the risk of contextual bias and the loss of information.

> The information that's in the statement might be biasing to you, or you see anonymous notes and they're full of vulgar language and things like that. That might be biasing. (QDE 15)

> You can't get away from that in circumstances that currently exist in our lab. So yes, there have been instances in our lab where I've been reading blood-stained notes and so on which could be a suicide note, or could be a threat or whatever it would be. I've done that. Very difficult to get away from that. (QDE 13)

> If you redact most of [the written words] off that page, it would be difficult to do the examination… I think it would make it harder to do the examination. (QDE 6)

> You'd just do your best to avoid that unconscious bias, but I think since it's part of the sample I don't think there's a way you can get around it, because you might have something important. Say if someone says "hey, I'm going to black this portion out first". You might have something really important that they have then taken out, and that changes the opinion of the case. (QDE 5)

Discussion

Most of our document examiners reported being aware of the potential for contextual bias in forensic science, and acknowledged that these biases could affect their work. These attitudes are similar to those of others in the forensic science community. In a global survey of forensic examiners' views on bias, the majority of respondents felt bias was a problem in the forensic sciences and that a forensic examiner's expectations can influence their analysis (Kukucka et al., 2017). Several examiners in Study 3, however, believed that knowledge of task-irrelevant contextual information did not put them at risk of being biased, either because they could ignore the information, or they would be sceptical towards the information. These opinions reflect arguments made by forensic examiners in other disciplines (Butt, 2013; Leadbetter, 2009; Thornton, 2010), but are not supported by empirical evidence (Dror et al., 2006; Dror & Charlton, 2006; Pronin, 2006).

When asked about task-relevance, our examiners' opinions were in agreement with the definitions provided by the National Commission on Forensic Science (NCFS, 2015). Our examiners reported that, more often than not, all they needed to know was the proposition in question. Contextual information was only considered task relevant if it helped the examiner attribute significance to features found in the evidence or if it informed their opinion level regarding authorship. That is, knowledge regarding factors that could have affected the physical appearance of the writing was the only type of task-relevant contextual information, and only on a case-by-case basis.

Understanding the task-relevance of contextual information is an important step in managing contextual bias, but equally important is determining *when* examiners should be made aware of this information. Research on context and comprehension indicates that disclosure of contextual information after the evaluation of the original information does not affect any subsequent interpretation of the original information (Bransford & Johnson, 1972). As such, document examiners should only be given the task-relevant information after the initial examination has been completed. The examiner can then re-examine the evidence (if necessary) and update their opinion. Any changes to the opinion should be recorded and taken into account by the trier-of-fact (e.g., jurors; Thompson, 2011).

Our document examiners considered contextual information management an important step towards mitigating bias due to task-irrelevant information. When asked about the contextual information management strategies they are currently using, several of our examiners reported successfully using strategies such as the case manager model or blind review. Others, however, reported experiencing difficulties in implementing any strategies due to resistance from co-workers, supervisors, or management. This resistance is likely due to the fact that many people "misconstrue cognitive bias as a motivational or ethical issue that can be overcome through sheer willpower, rather than an intrinsic feature of human nature" (Kukucka et al., 2017, p. 456).

Interestingly, one document examiner's approach to peer review was to try and prove the initial examiner wrong; the verifier essentially approaches their task with a *consider-the-opposite* strategy, which can mitigate the risk of confirmation bias (Reese, 2012). Confirmation bias involves a positive test strategy, where a person will test a hypothesis by searching for, or prioritizing, data that confirms rather than confutes it (Saks et al., 2003; Wason, 1960). The purpose of the consider-the-opposite strategy is to reduce biases by encouraging the verifier to focus on features and characteristics that violate their expectations. This approach to hypothesis testing has been shown to be a successful debiasing technique in several different domains. For example, Budesheim and Lundquist (1999) discovered that arguing for a position inconsistent with one's beliefs reduced participants' tendency to consider information in a biased manner, and Hirt and Markman (1995) found that participants formed debiased judgments of hypothetical situations when they considered alternative or opposite explanations for the situation. The consider-the-opposite strategy is currently not commonplace in the verification of conclusions in forensic disciplines. Considering that this strategy has been successful in reducing confirmation bias in other domains, it could be an effective method of reducing bias in forensic examiners.

Unfortunately, the contextual information management strategies described by our document examiners did not protect them from external pressure; some of our examiners described experiences in which they were pressured by their clients or factors in their work environment to reach a specific conclusion. Reports of clients pressuring forensic examiners appear to be relatively common in the literature. According to Giannelli (2007, p. 226), attorneys often "push the envelope" in pressuring their experts to give the desired opinion. A survey of 365 forensic examiners in the UK revealed that 28.6% of respondents experienced feeling pressured to produce a particular result (Geddes, 2012).

Indeed, there are documented cases in which investigator pressure on forensic examiners resulted in a wrongful conviction. In 2010, the North Carolina state crime laboratory came under investigation when laboratory results led to the wrongful conviction of a man named George Taylor (Giannelli, 2012). Investigators discovered that the crime laboratory had distorted or withheld evidence in more than 230 cases since 1994. According to the investigators, this misconduct was due to forensic examiners coming under pressure from prosecutors to produce results that would secure a conviction. Most notably, the laboratory had a policy in place preventing the reporting of results of more sophisticated tests if they challenged or contradicted those of simpler tests that supported the prosecution (Giannelli, 2012). Cases such as these demonstrate the importance of forensic examiners working with laboratory management to secure proper contextual information management strategies which protect them from external pressure.

During our interviews, however, it became clear that external pressure is not the only barrier document examiners face in managing contextual bias. One major barrier to implementing contextual information management strategies, for example, appeared to be concern about the fact that there is uncertainty regarding how effective suggested contextual information management strategies are. In forensic science, the fact that the ground truth is often impossible to determine makes it challenging to determine whether the implementation of context management is having any effect (Park, 2008). Furthermore, both private and government laboratories have limited resources, most of which are dedicated to efficient turnaround times. As a result, management can be reluctant to spend valuable resources in the absence of feedback to support the efficacy of contextual information management strategies (but see Chapter 9, for a discussion on inserting blind-tests into the workflow).

Where to from here?

Almost 125 years have passed since Hagan (1894) first raised concerns about bias in handwriting examination. The interviews conducted in Study 3 have made it clear that these concerns are still not being adequately addressed by modern-day document examiners. Although some practitioners have taken steps to mitigate the risk of bias, the implementation of contextual information management strategies is occurring at a slow pace, and is being met with resistance from those who still believe that bias is not the problem it is being made out to be. In Chapter 9, we reflect on this issue in light of our experimental and field data, and consider the way forward.

CHAPTER NINE

GENERAL DISCUSSION

Although people are inclined to believe that they perceive and interpret the world around them in an objective and unbiased way, the reality is often quite different (Pronin et al., 2002). Instead, when faced with ambiguous or incomplete information, we rely on top-down processes to ‘fill in the gaps’. These processes place us at risk of biased interpretations of the information, which can lead to error (Bartels, 2010; Fraser-Mackenzie et al., 2013; Giannelli, 2007, Kinchla & Wolfe, 1979). Of course, in everyday life, such errors usually only have minor consequences—we might misremember an argument with a friend in a way that favours our own view, for example. Unfortunately, however, there are numerous contexts in which bias can have severe ramifications. Biased interpretations of forensic evidence, for example, have been associated with numerous wrongful convictions (Garrett & Neufeld, 2009; Innocence Project, 2017a; 2017b; 2017c; National Registry of Exonerations, 2018). In this thesis, we explored the effect of contextual bias in handwriting analysis—a forensic discipline that has received little research attention, despite several characteristics that render its examiners particularly susceptible to context-induced bias (Kukucka & Kassin, 2014; Miller, 1984; Sulner, 2014).

Handwriting differs from other types of forensic evidence, chiefly due to its highly dynamic nature. That is, a considerable level of intra-writer variation means that differences between a questioned sample of handwriting and a known sample cannot necessarily be attributed to a different source. To date, however, only two studies have directly investigated the effects of contextual information on handwriting analysis (Kukucka & Kassin, 2014; Miller, 1984). Furthermore, only one unpublished study has explored contextual bias as it relates to signature examination (reported in Merlino, 2015)—in which examiners need to base their decisions on very small writing samples. The overarching goal of this thesis was to address this gap in the literature.

Experimental Research

In Studies 1A and 1B, lay participants were presented with a series of questioned and known signatures, and asked to determine whether or not the questioned signature had been forged. Half of the questioned signatures were accompanied by high-biasing case information suggesting a forgery (e.g., stating that the questioned signature came from a suspect with a criminal record), the other half were accompanied by low-biasing information (i.e., simply stating that the questioned signature came from a suspect). Contrary to our expectations, the contextual manipulation did not significantly influence participants' ability to discriminate between genuine signatures, or their response bias. We suggested two possible reasons why we did not observe a context effect in Study 1A. The first was that the response scale we used was not sensitive enough—the use of a dichotomous forged or genuine response option may have obscured any context effects on the examination process. The second possibility was that the contextual information was not directly tied to the decision-making process—participants read the information prior to being exposed to the signatures.

We controlled for these two possibilities in Study 1B. Contrary to Study 1A, signal detection analyses in Study 1B revealed that the biasing information affected both participants' sensitivity and response bias. The presence of highly-biasing contextual information significantly reduced participants' ability to discriminate forged from genuine signatures, and increased their tendency to say that signatures were forged. Furthermore, although the contextual manipulation did not influence participants' confidence in their 'forged' decisions, it significantly reduced their confidence in their 'genuine' decisions.

The presence of a context effect on sensitivity in Study 1B raised the possibility that the contextual information may have influenced the process by which participants' examined questioned signatures. Specifically, directional cues in the contextual information could have led our participants to disproportionately focus on features in the evidence that supported the expected outcome (i.e., to engage in a positive test strategy). In Study 2, we therefore investigated the association between contextual bias and the signature examination process. More specifically, we adopted the same paradigm used in Experiment 1B, but manipulated whether people searched for and identified similarities or differences between questioned and known signatures. If contextual information promotes the use of a positive test strategy (and thus, contextual bias), then forcing participants to engage in a negative test strategy should eliminate the effect.

Contrary to our expectations, instructing participants to focus on similarities did not eliminate contextual bias. In fact, participants who focused on similarities exhibited a significantly *higher* response bias (that is, a *lower* threshold for determining that the signature was forged), relative to participants who were asked to identify differences. Our Study 2 data also showed that, regardless of task, the features participants searched for and identified in the signatures was firmly tied to the contextual information. Specifically, participants who were asked to identify similarities indicated more points in low-bias trials than in high-bias trials, while those asked to identify differences showed the opposite pattern.

Taken together, our data show some evidence of contextual bias in a signature examination task. These data are also broadly consistent with the existing literature on confirmation bias (Giannelli, 2007; Klayman & Ha, 1987; Nickerson, 1998; Wason, 1960), and contribute to a growing body of research demonstrating that context-driven expectations can affect the evaluation and judgment of forensic evidence (e.g., Dror et al., 2006; Dror & Charlton, 2006; Kukucka & Kassin, 2014; Lange et al., 2011; Langenburg et al., 2009).

Our signal detection data are also consistent with those of previous studies that have found that laypeople show a liberal response bias—that is, a higher rate of hits and false alarms than misses and correct rejections—regardless of their exposure to contextual information (Searston et al., 2015; Stevenage & Bennett, 2017). In our discussions of Study

1 and 2, we suggested that the mere fact that the signature was in question could have led participants to expect it to be forged, and therefore that the suspect was guilty.

In handwriting examinations more generally, however, deciding that two samples were written by different people does not always indicate guilt. Rather, this notion is dependent on the proposition associated with the questioned writing. Consider, for example, the contextual information used by Miller (1987) and Kukucka and Kassin (2014). Participants in those studies were required to determine whether a suspect had written a hold-up note; in this case, a match indicated guilt. We therefore propose a follow-up study to Study 1 and Study 2, in which these propositions are manipulated in a within-subjects fashion—that is, in some trials, a match would indicate guilt, while in others, a non-match would indicate guilt. Coupled with contextual information suggesting either a match or a non-match, such a study would provide more nuanced insight into the effects of contextual information on evaluations of questioned handwriting.

One distinct advantage of the signal detection approach that we took is that it allowed us to separate the effects of contextual information on sensitivity and response bias. A limitation of this framework, however, is that it necessitated a dichotomous response. Such a response scale is not representative of that used by questioned document examiners, who are able to provide nine possible conclusions, ranging from *identification* (i.e., "John Doe wrote the questioned material") to *inconclusive* (i.e., "I am unable to determine whether John Doe wrote the questioned material") to *elimination* (i.e., "John Doe did not write the questioned material") (SWGDOC, 2013). Under such circumstances, it may be more appropriate to utilise a rating task in which participants give graded responses (e.g., 1 = identification, 9 = elimination) instead. While we took this approach in Study 2, it only allowed us to examine participants' confidence independently, rather than factoring it into our signal detection analyses.

According to Stanlisaw and Todorov (1999), continuous ratings can be used to calculate points on a *receiver operating characteristic* (ROC) curve, which plots hit rate as a function of false alarm rate. Each point on the curve corresponds to a different criterion; for example, one point distinguishes ratings of 1 from ratings of 2, another distinguishes ratings of 2 from ratings of 3, and so on. The area under the curve is a measure of sensitivity, independent from response bias. An ROC approach does not, however, offer a direct measure of overall response bias. Instead, it provides an index of decision thresholds—each point on the curve represents a subjective threshold point for each comparison (Phillips et al., 2001). Future research should begin to explore the use of ROC analysis in contextual bias studies.

Another important aspect of Studies 1 and 2 that warrants attention is the nature of our sample. Specifically, the fact that our participants were laypersons and not forensic examiners could lead some to argue that our findings do not inform professional practice (Langenburg, 2017). Forensic examiners are considered experts in their respective fields on the basis that they can perform their tasks to a higher standard than novices (Towler et al., 2018). Studies comparing forensic examiners and novices have repeatedly shown that forensic experts consistently outperform novices in terms of accuracy (Bird et al., 2010b; Busey et al., 2010; Dyer et al., 2006; Kam et al., 1997; Kam et al., 2001; Kam et al., 2015; Searston & Tangen, 2017; Sita et al., 2002; Tangen et al., 2011; Thompson & Tangen, 2014; Thompson, Tangen, & McCarthy, 2014; Towler, White, & Kemp, 2017), leading some to question the appropriateness of extrapolating findings from studies using laypeople to the wider forensic community (Langenburg, 2017).

It is important to note, however, that our participants' mean sensitivity (*d'*) scores were significantly higher than zero. That is, even though they were not trained in signature examination, our participants were able to discriminate between genuine and forged signatures at a rate that was significantly greater than chance. This finding suggests that we can generalise our data to some extent, in a way that we could not if participants' responses were essentially random (i.e., a mean sensitivity that did not differ significantly from zero).

It is sometimes argued that forensic examiners' expertise should make them less susceptible to contextual bias (Langenburg, 2017). When we consider the wider literature on the nature of expertise, however, we see that this is unlikely to be the case. In fact, some researchers have argued that experts are likely to be even *more* vulnerable to bias than laypeople (Dror & Charlton, 2006). Expertise entails the use of schemas, heuristics, and top-down processing to perform tasks quickly and efficiently. Unfortunately, reliance on these cognitive processes renders experts more vulnerable to tunnel vision and bias than laypeople (Busey & Dror, 2011; Dror, 2011). Forensic examiners might also be especially susceptible to bias because of the emotional nature of their work, which could promote motivational bias—especially in examiners who exhibit a high need for cognitive closure (Charlton et al., 2010).

Another limitation of Studies 1 and 2 with regard to ecological validity is that our paradigm and stimuli were not a precise analogue of the everyday work environment of a handwriting examiner. They did, however, afford us a high degree of experimental control and a level of statistical power that is not available in field research using very highly specialised—and therefore small—populations. Experimental studies carried out in forensic science workplaces offer ecological validity, but often involve very small samples (e.g. Dror & Charlton, 2006; Dror et al., 2006; Dror & Hampikian, 2011; Hall & Player, 2008; Kerstholt et al., 2007; Kerstholt et al., 2010; Langenburg et al., 2009). Furthermore, in several field studies of contextual bias in forensic disciplines, the experts were aware they were taking part in a research study (e.g., Dror & Hampikian, 2011; Hall & Player, 2008; Langenburg, et al., 2009), raising the risk that they would complete the tasks in a way that did not reflect their usual practice (Dror, 2009; Dror & Cole, 2011; Dror & Rosenthal, 2008; Park, 2008; see the

Hawthorne effect; Adair, 1984; Levitt & List, 2011). Studying expert performance becomes difficult because participants are motivated to do well—their performance during the experiment is not going to be representative of their routine day-to-day casework (Dror & Rosenthal, 2008; Park, 2008).

To overcome this problem, Dror et al. (2006) and Dror and Cole (2006) covertly inserted test cases into the daily workflow of expert fingerprint examiners. Not only did these researchers find evidence to suggest that fingerprint examiners could be biased by contextual information, they also discovered that their participants' expertise did not protect them from such effects. Unfortunately, although these two studies demonstrated a high degree of ecological validity, they were limited by very small samples (N = 5, and N = 6, respectively).

In conclusion, although our two lab-based studies allowed us to maintain a high degree of experimental control over the stimuli and manipulated variables, it is clear that investigating expert performance in an ecologically valid way while maintaining a degree of scientific rigor poses a considerable challenge. As such, we argue that the findings from the two experimental studies should not be taken as evidence for contextual bias in real-world signature examinations, but rather be seen as a first step towards developing signal detection paradigms that can be used to investigate context effects in professional handwriting examiners.

Field Research

In addition to calling for studies that investigate contextual bias in forensic examiners, the NRC (2009) recommended that forensic laboratories develop standard operating procedures that minimize or mitigate the risk of bias. In light of our Study 2 data—which suggest that simply manipulating the way the examiners search for information is unlikely to combat contextual bias—it is particularly crucial to examine potential ways to either limit exposure to contextual information, or to find ways to present contextual information in a way that is least likely to influence examiners' decisions.

Academic commentators have proposed several strategies for managing contextual information in forensic science disciplines (Budlowe et al., 2009; Dror et al., 2015; Found & Ganas, 2013; Kassin et al., 2013; Krane et al., 2008; Saks et al., 2003; Stoel et al., 2011; Thompson, 2011; Wells et al., 2003). The most extreme of these recommendations involve blinding forensic scientists to *all* contextual information (PCAST, 2016). Oftentimes, however, forensic examiners need a degree of contextual information in order to perform the task asked of them (Thompson, 2011; Thornton, 2010). In these cases, what becomes crucial is an understanding of which contextual information is—and is not—relevant. Because handwriting is the product of both behaviour and circumstances (e.g., the writing surface, or the age of the writer), the type of contextual information that a handwriting examiner requires is likely to differ from that required in other forensic disciplines. Some or all of the recommended context management strategies might, therefore, not be feasible for document examination.

In Study 3, we interviewed professional document examiners about their view on—and experiences with—contextual bias in their daily work. Our goals were to investigate their perceptions of the relevance of contextual information in document examination, the current context management strategies in place at their workplaces, and the barriers to implementing contextual information management strategies.

Our interviewees reported being aware of the potential for contextual bias in forensic science, and considered contextual information management an important step towards mitigating bias that could arise due to task-irrelevant information. Unfortunately, however, the majority of our interviewees reported institutional or financial barriers that prevented them from implementing most or all of the context management recommendations made by academic commentators. One common theme discussed by the document examiners was that many of the recommendations may not be practical or feasible in their line of work. Because the NRC (2009) report recommended that forensic laboratories develop standard operating procedures that minimize or mitigate the risk of bias, we will discuss contextual information management strategies that are likely to be effective for document examination in light of our findings as a whole.

The results of our experimental studies indicate that knowledge of contextual information can reduce sensitivity and increase response bias, and that contextual bias cannot be eliminated by asking people to focus on features in the evidence that disconfirm directional cues in the contextual information. As such, it is imperative to minimize forensic examiners' exposure to contextual information. One proposed strategy to achieve this is the use of a case manager in forensic laboratories.

Several document examiners in Study 3 reported working in an environment that had successfully introduced case managers. Other examiners, however, questioned the practicality of having a nominated case manager in their laboratory. Their primary concern was lack of knowledge on the case manager's behalf—document examiners believed that, for a case manager to be effective, they would have to have an understanding of how document examination worked and what information was relevant to the process. Because it is not realistic to expect a case manager to be trained in all forensic disciplines that they would be managing cases for, we suggest a compromise. The document examiners interviewed in Study 3 identified several types of information that could be task-relevant, and largely agreed on the information that is entirely irrelevant. We suggest that similar interviews are conducted with practitioners in other forensic disciplines, so a list of task-relevant information can be developed. Such a list would aid a case manager in determining which information to remove from a case, and qualified forensic examiners would be able to focus on their work rather than divide their time between two roles.

Some questioned document laboratories—those involving sole-practitioners, for example—do not have the resources to establish a case manager in their laboratory. A resource-efficient alternative to a case manager is linear sequential unmasking, in which the questioned writing is examined first in the absence of the known (suspect) material (Huber & Headrick, 1999; Lewis, 2014; SWGDOC, 2000). Such a strategy prevents examiners from "working backwards" from the suspect to the crime scene and (knowingly or unknowingly) allowing their expectations to influence the examination process (Dror, 2013). This approach might be especially important considering findings from our experimental studies suggest that context-driven expectations and directional cues inherent in contextual information negatively affected our participants' sensitivity and response bias.

An additional step raised by the practitioners we interviewed was making the verification part of the ACE method a blind one (Budlowe et al., 2009; Dror et al., 2015; Osborne & Taylor, 2018). That is, a second examiner would re-work the case without awareness of (1) any case-related information and (2) the first examiner's conclusions. Although many of the document examiners interviewed in Study 3 reported that the verification or technical review of a case is common practice in their workplace, this review process is often not blind; instead, the reviewer merely examines the original examiner's notes to assess whether the conclusions are supported by the evidence. This type of verification has been roundly criticised in other forensic disciplines, as knowledge of the first examiner's opinion can influence the verifier's conclusions (Langenburg, Champod, & Wertheim, 2009; Merlino, 2015; OIG, 2006). A similar effect was also demonstrated by participants in Study 2: knowledge of forensic examiners' opinions influenced our participants' evaluations of the evidence.

Notably, one of our interviewees reported trying to prove the initial examiner wrong during verification (*consider-the-opposite* strategy; Reese, 2012). Adopting such a strategy could be an alternative to blind reviews for laboratories that are short on time and personnel, but caution is warranted considering that such a strategy likely created a debiasing backfire effect in participants in Study 2 (Sanna et al., 2002; Schwarz & Vaughan, 2002). That is, asking participants to focus on features in the evidence that contradicted the suggestion that the questioned signature was forged resulted in those participants being *more* biased towards concluding a forgery.

One concern expressed by the practitioners that we interviewed was a lack of empirical evidence on the effectiveness of contextual information management strategies. As one of our interviewees stated, “We’ve got no measure of how effective this is down the track, like whether it [context management] has made any difference or not” (QDE 16, p. 6). Another interviewee commented that “I would be interested in finding out more about the balance between how much influence does it [contextual information] potentially have, so how serious is the problem versus how much impact would any mitigating management that we try and infuse it to have. Also, would that management actually be real and would we be actually achieving anything other than giving the appearance that we’re doing anything about it” (QDE 17, p. 4). It is challenging to measure the success of context management schemes because the ground-truth in everyday cases is unknown.

One way of testing the efficacy of context management methods would be to integrate blind-tests into the workflow, where the ground truth is known and error rates can be calculated (Whitman & Koppl, 2010). In the forensic literature, double blind testing refers to tests in which an examiner does not know the ground truth and is not aware that they are being tested. Kerkhoff et al. (2015), for example, trialled a double-blind testing programme in firearms examination in the Netherlands. The authors prepared proficiency tests and validity studies and asked police agencies to submit these as real cases. Not only did this allow the authors to calculate the rate of incorrect conclusions, knowing the ground truth meant that results from these tests could be used to provide examiners with feedback on their performance. It should be noted that Kerkhoff et al. (2015) used the double-blind test to investigate error rates, rather than bias specifically. However, a similar methodology could be applied to investigate the efficacy of context management strategies.

Concluding Remarks

Although there are still individuals in the forensic community have difficulty recognizing the fact that contextual bias can influence the examination of forensic evidence (e.g., Butt, 2013; Leadbetter, 2007; Thornton, 2010), the majority of the forensic community has now accepted that contextual bias poses a real threat to their objectivity. Indeed, a large body of research has demonstrated that contextual bias can negatively affect the examination and interpretation of forensic evidence (Dror et al., 2005; Dror et al., 2006; Dror & Charlton, 2006; Kassin et al., 2013; Kukucka & Kassin, 2014; Langenburg et al., 2009; Nakhaeizadeh, et al., 2014; Osborne & Zajac, 2015; Osborne et al., 2016a; Searston et al., 2015; Stevenage & Bennett, 2017), and members of the forensic and academic community have established working groups dedicated to investigating this very issue (e.g., the Organization of Scientific Area Committees (OSAC) for Forensic Science).

The next step should be to develop and test standardized guidelines for contextual information management that take into account the different needs of sole-, private, and government document examination laboratories. The challenges faced by document examiners in the private sector are different from those who work in a government laboratory. It is clear that current recommendations for context management cannot take a 'one size fits all' approach, but this has not been addressed in proposed contextual information management strategies, and a large portion of document examiners are currently not able to implement the suggested contextual information management strategies.

Furthermore, document examiners should not be held solely responsible for the management of contextual information. From our interviews it is clear that attorneys, clients, and investigators are a significant source of contextual information and bias, and yet context management recommendations to date have not addressed this (Elaad, 2013). We recommend the development of guidelines and training programs to educate investigators on the dangers of exposing document examiners to case-related information. It would also be beneficial to implement institute-wide policies that protect the examiners from being exposed to extraneous information. Furthermore, in the civil arena, we suggest that attorneys be educated on the dangers of biasing their experts, and that legal support is provided for document examiners who feel they are being biased or pressured by their retaining attorney.

Finally, the efficacy of context management strategies needs to be tested. We suggest inserting blind tests—using a signal detection paradigm—into the daily workflow. Although this might not be feasible for all laboratories, especially those involving sole practitioners, we argue that tests of this nature can carried out in larger laboratories to establish baseline efficacy. Once the efficacy of context management has been tested, and if these methods are found to be effective, academic commentators and forensic scientists could collaborate to develop standardized contextual management strategies for each forensic discipline. Their overarching aim to should be to establish standardized guidelines on context management across all forensic laboratories.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of a methodological artefact. *Journal of Applied Psychology*, *69,* 334-345.

Amato, P. (1991). The "child of divorce" as a person protoype: Bias in the recall of information about children in divorced families. *Journal of Marriage and Family*, *53,* 59-69.

American Board of Forensic Document Examiners. (2014). Rules and procedures guide. Retrieved from https://www.abfde.org/htdocs/certification/QualificationsandRequirementsforCerts.pdf.

Andersen, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Education Research Journal, 14,* 367-381.

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of Investigative Psychology and Offender Profiling*, *2,* 43-63.

Ask, K., Rebelius, A., & Granhag, P. (2008). The 'elasticity' of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology*, *22,* 1245-1259.

Baker, G. R., Norton, P. G., Flintoft, V., Blais, R., Brown, A., Cox, J., Etchells, E., Ghali, W. A., Hébert, P., Majumdar, S. R., O'Beirne, M., Palacios-Derflingher, L., Reid, R. J., Sheps, S., & Tamblyn, R. (2004). The Canadian adverse events study: The incidence of adverse events among hospital patients in Canada. *Canadian Medical Association Journal*, *170,* 1678-1686.

Ballantyne, K. N., Edmond, G., & Found, B. (2017). Peer review in forensic science. *Forensic Science International*, *277,* 66-76.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, *42*, 88-110.

Bartels, B. (2010). Top-down and bottom-up models of judicial reasoning. In D. E. Klein & G. Mitchell (Eds.), *The psychology of judicial decision making* (pp. 41-56). Oxford University Press.

Bird, C., Found, B., Ballantyne, K., & Rogers, D. (2010a). Forensic handwriting examiners' opinions on the process of production of disguised and simulated signatures. *Forensic Science International*, *195*, 103-107. doi: 10.1016/j.forsciint.2009.12.001

Bird, C., Found, B., & Rogers, D. (2010b). Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviours. *Journal of Forensic Sciences*, *55*, 1291-1295. doi: 10.1111/j/1556-4029.2010.01456.x

Bobrow, D. G., & Norman, D. A. (1975). Some principles of memory schemata. In D. G. Bobrow & A Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 131-149). New York: Academic Press.

Bransford, J., & Johnson, M. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717-729.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77-101.

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

Budesheim, T. L., & Lundquist, A. R. (1999). Consider the opposite: Opening minds through in-class debates on course-related controversies. *Teaching of Psychology*, *26*, 106-110.

Budlowe, B. (2010). A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement [Author's response]. *Journal of Forensic Sciences*, *55*, 275-276.

Budlowe, B., Bottrell, M. C., Bunch, S. G., Fram, R., Harrison, D., Meagher, S., Oien, C. T., Peterson, P. E., Seiger, D. P., Smith, M. B., Smrz, M. A., Soltis, G. L., & Stacey, R. B. (2009). A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *Journal of Forensic Sciences*, *54*, 798-809.

Busey, T., & Dror, I. E. (2011). Special abilities and vulnerabilities in forensic expertise. In *The Fingerprint Sourcebook*. NIJ Press: Washington, DC.

Busey, T., Yu, C., Wyatte, D., Vanderkolk, J., Parada, F., & Akavipat, R. (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, *61*, 60-91.

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

Byers, S. N. (2016). *Introduction to Forensic Anthropology*. New York: Taylor & Francis.

Cappelli, R., Ferrara, M., & Maltoni, D. (2012). Minutiae-based fingerprint matching. In C. Liu & V. K. Mago (Eds.), *Cross Disciplinary Biometric Systems* (pp.117-150). doi: 10.1007/97836422845717

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology*, *7*, 91-103.

Centers for Disease Control and Prevention. (2018). *WONDER database*. Retrieved from https://wonder.cdc.gov/.

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behaviour. *Journal of Personality and Social Psychology*, *50*, 492-501.

Charlton, D., Fraser-Mackenzie, P. A. F., & Dror, I. E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. *Journal of Forensic Sciences*, *55*, 385-393.

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.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting the truth or getting along: Accuracy-versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*, 262-275.

Cole, S. A. (2016). Scandal, fraud, and the reform of forensic science: The case of fingerprint analysis. *West Virginia Law Review*, *119*, 523-548.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1-73.

Davis, D., Lopez, P., Koyama, M., White, B., Knaack, D., Kusal, T., et al. (2005). *Memory for threats in conversation enhanced by later knowledge of violence between participants*. Paper presented at the American Psychological Society, Los Angeles, CA.

Department of Justice, Office of Public Affairs. (2017, April 10). *Attorney General Jeff Sessions announces new initiatives to advance forensic science and help counter the rise in violent crime* [Press release]. Retrieved from https://www.justice.gov/opa/pr/attorney-general-jeff-sessions-announces-new-initiatives-advance-forensic-science-and-help

Devine, D. J., Clayton, , L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, *7*, 622-727.

DiFonzo, J. H., & Stern, R. C. (2007). Devil in a white coat: The temptation of forensic evidence in the age of CSI. *New England Law Review*, *41*, 503-532.

Dooling, D. J., & Lachman, R. (1971). Effects of comprehension on retention of prose. *Journal of Experimental Psychology*, *88*, 216-222.

Dror, I. E. (2009). On proper research and understanding of the interplay between bias and decision outcomes [Letter to the editor]. *Forensic Science International*, *191*, 17-18.

Dror, I. E. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.), *The paradoxical brain* (pp. 177-188). Cambridge University Press.

Dror, I. E. (2013). Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management: An International Journal*, *4*, 3-4.

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.

Dror, I. E., & Cole, S. (2010). The vision in “blind” justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*, 161-167.

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science and Justice*, *51*, 204-208.

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

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.

Dror, I., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and bias ability of forensic experts. *Journal of Forensic Sciences*, *53*, 900-903.

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making [Letter to the editor]. *Journal of Forensic Sciences*, 1-2.

Dyer, A., Found, B., & Rogers, D. (2006). Visual attention and expertise for forensic signature analysis. *Journal of Forensic Sciences*, *51*, 1397-1404. doi: 10.1111/j.1556-4029.2006.00269x

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R. A., Thompson, M. B., Tangen, J. M., Kemp, R. I., & Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science and Justice*, *57*, 144-154.

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

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*, 311-318.

Epley, N., Keysar, B., van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87,* 327-339.

Evans, J. S. B. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*, 313-326.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher-cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223-241.

Evett, I. W., Berger, C. E. H., Buckleton, J. S., Champod, C., & Jackson, G. (2017). Finding the way forward for forensic science in the US—A commentary on the PCAST report. *Forensic Science International*, *278,* 16-23.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behaviour: The mode model as an integrative framework. *Advances in Experimental Social Psychology*, *23*, 75-109.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 97-116). New York, NY: Guilford Press.

Federal Bureau of Investigation. (n.d.). *Lindbergh Kidnapping*. Retrieved from https://www.fbi.gov/history/famous-cases/lindbergh-kidnapping.

Federal Bureau of Investigation. (2015). *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review* [Press release]. Retrieved from https://www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review.

Fischer, P., Jonas, E., Frey, D., & Schulz-Hardt, S. (2005). Selective exposure to information: The impact of information limits. *European Journal of Social Psychology*, *35,* 469-492.

Fiske, S. T., & Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin*, *6*, 543-557.

Found, B., & Ganas, J. (2013). The management of domain irrelevant context information in forensic handwriting examination casework. *Science and Justice*, *53*, 154-158. doi: 10.1016/j.scijus.2012.10.004

Found, B., & Rogers, D. (2008). The probative character of forensic handwriting examiners' identification and elimination opinions on questioned signatures. *Forensic Science International*, *178*, 54-60. doi: 10.1016/j.forsciint.2008.02.001

Frantz, C. M. P. (2010). I AM being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology*, *28*, 157-167.

Fraser-Mackenzie, P. A. F., Bucht, R. E., & Dror, I. E. (2013). Forensic judgment and decision making. In T. R. Zentall & P. H. Crowley (Eds.), *Comparative decision making* (pp. 386-417). New York, NY: Oxford University Press.

Frey, D., Schulz-Hardt, S., & Stahlberg, D. (1996). Information seeking among individuals and groups and possible consequences for decision-making in business and politics. In E. Witte & J. Davis (Eds.), *Understanding group behaviour: Vol II. Small group processes and interpersonal relationships* (pp. 211-225). Mahwah, NJ: Erlbaum.

Gallimore, P. (1996). Confirmation bias in the valuation process: A test for corroborating evidence. *Journal of Property Research*, *13*, 261-273.

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

Geddes, L. (2012). *Forensic failure: ‘Miscarriages of justice will occur’*. Retrieved from https://www.newscientist.com/article/mg21328514-600-forensic-failure-miscarriages-of-justice-will-occur/#bx285146B1

Giannelli, P. C. (2007). Wrongful convictions and forensic science: The need to regulate crime labs. *North Carolina Law Review*, *86*, 163-235.

Giannelli, P. C. (2012). The North Carolina crime lab scandal. *Criminal Justice*, *27*, 43-45, 62-63.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases”. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (pp. 83-115). Chichester, UK: Wiley.

Giles, A. (2004). The forensic examination of documents. In P. White (Ed.), *Crime Scene to Court: Essentials of Forensic Science* (2nd ed., pp. 142-171). Cambridge: The Royal Society of Chemistry.

Gilovich, T., & Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-18). New York, NY: Cambridge University Press.

Gleitman, H., Gross, J., & Reisberg, D. (2011). *Psychology*. New York, NY: W. W. Norton & Company, Inc.

Grether, D. M. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, *17,* 31-57.

Haber, R. N., & Haber, L. (2013). The culture of science: Bias and forensic evidence. *Journal of Applied Research in Memory and Cognition*, *2*, 65-67.

Hagan, W. E. (1894). *A treatise on disputed handwriting and the determination of genuine from forged signatures*. New York, NY: Banks & Brothers.

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.

Halverson, A. M., Hallahan, M., Hart, A. J., & Rosenthal, R. (1997). Reducing the biasing effects of judges' nonverbal behaviour with simplified jury instruction. *Journal of Applied Psychology, 82*, 590-598.

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

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.

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-the-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69,* 1069-1086.

Huber, R. A., & Headrick, A. M. (1999). *Handwriting identification: Facts and fundamentals*. Boca Raton: CRC Press.

Innocence Project. (2017a). *The Cases: Kennedy Brewer*. [online] Retrieved from https://www.innocenceproject.org/cases/kennedy-brewer/

Innocence Project. (2017b). *The Cases: Larry Fuller*. [online] Retrieved from https://www.innocenceproject.org/cases/larry-fuller/

Innocence Project. (2017c). *The Cases; William Barnhouse*. [online] Retrieved from https://www.innocenceproject.org/cases/william-barnhouse/

Innocence Project. (2018). *DNA Exonerations in the United States*. [online] Retrieved from https://www.innocenceproject.org/dna-exonerations-in-the-united-states/

Jackson, J. C., Jong, J., Bluemke, M., Poulter, P., Morgenroth, L., & Halberstadt, J. (2017). Testing the causal relationship between religious belief and death anxiety. *Religion, Brain & Behavior, 8*, 57-68. doi: 10.1080/2153599X.2016.1238842

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161-1166.

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

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

Kam, M., Abichandani, P., & Hewett, T. (2015). Simulation detection in handwritten documents by forensic document examiners. *Journal of Forensic Sciences, 60*, 936-941. doi: 10.1111/1556-4029.12801

Kam, M., Fielding, G., & Conn, R. (1997). Writer identification by professional document examiners. *Journal of Forensic Sciences, 42,* 1-33.

Kam, M., Gummadidala, K., Fielding, G., & Conn, R. (2001). Signature authentication by forensic document examiners. *Journal of Forensic Sciences, 46*, 884-888.

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

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

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

Kerkhoff, W., Stoel, R. D., Berger, C. E., H., Mattijssen, E. J. A. T., Hermsen, R., Smits, N., & Hardy, H. J. J. (2015). Design and results of an exploratory double blind testing program in firearms examination. *Science and Justice, 55*, 514-519.

Kerstholt, J. H., Paashuis, R., & Sjerps, M. (2007). Shoe print examinations: Effects of expectation, complexity, and experience. *Forensic Science International*, *165,* 30-34.

Kerstholt, J., Eikelboom, A., Dijkman, T., Stoel, R., Hermsen, R., & van Leuven, B. (2010). Does suggestive information cause a confirmation bias in bullet comparisons? *Forensic Science International*, *198*, 138-142.

Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing: "Top-down," "bottom-up," or "middle out". *Perception & Psychophysics*, *25,* 226-231.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385-418.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Kleider, H. M., Pezdek, K., Goldinger, S. D., & Kirk, A. (2008). Schema-driven source misattribution errors: Remembering the expected from a witnessed event. *Applied Cognitive Psychology, 22,* 1-20.

Knoblish, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29,* 1000-1009.

Koppenhaver, K. (2007). *Forensic document examination: Principles and practice*. Totowa, NJ: Humana Press.

Krane, D. E., Ford, S., Gilder, J. R., Inman, K., Jamieson, A., Koppl, R., Kornfield, I. L., Risinger, D. M., Rudin, N., Taylor, M. S., & Thompson, W. C. (2008, July). Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation [Letter to the editor]. *Journal of Forensic Science*, *53,* 1006-1007.

Kray, L. J., & Galinsky, A. D. (2003). The debiasing effect of counterfactual mind-sets: Increasing the search for disconfirmatory information in group decisions. *Organizational Behavior and Human Decision Processes*, *91*, 69-81.

Kruger, J., & Gilovich, T. (1999). “Naïve cynicism” in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, *76*, 743-753.

Kukucka, J. (2014). The journey or the destination? Disentangling process and outcome in forensic identification. *Forensic Science Policy & Management*, *5*, 112-114. doi: 10.1080/19409044.2014.966928

Kukucka, J., & Kassin, S. (2014). Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias. *Law and Human Behavior*, *38*, 256-270. doi: 10.1037/lhb0000066

Kukucka, J., Kassin, S. M., Zapf, P. A., & Dror, I. E. (2017). Cognitive bias and blindness: A global survey of forensic science examiners. *Journal of Applied Research in Memory and Cognition*, *6*, 452-459.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480-498.

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

Langenburg, G. (2017). Addressing potential observer effects in forensic science: A perspective from a forensic scientist who uses linear sequential unmasking techniques. *Australian Journal of Forensic Sciences*, *9*, 548-563.

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.

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

Leape, L. L., Berwick, D. M., & Bates, D. W. (2002). Counting deaths due to medical errors—Reply. [Letter to the editor]. *JAMA*, *288*, 2405.

Leblanc, V. R., Brook, L. R., & Norman, G. R. (2002). Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. *Academic Medicine*, *77*, 67-69.

Levitt, S. D., & List, J. A. (2011).Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics*, *3*, 224-238.

Lewis, J. (2014) *Forensic document examination: Fundamentals and current trends*. Burlington: Elsevier Science.

Lewis, J. A. (2016). Minimizing cognitive bias in forensic document examination. *Journal of the American Society of Questioned Document Examiners, Inc.*, *19*, 33-36.

Lieberman, J. D., Carrell, C. A., Miethe, T. D., & Krauss, D. A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law*, *14*, 27-62.

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.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585-589.

Loftus, E. F., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society*, *5*, 86-88.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory & Psychology*, *1*, 65-82.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231-1243.

Luftig, M. A., & Richey, S. (2001). DNA and forensic science. *New England Law Review*, *35*, 609-613.

Lundgren, S. R., & Prislin, R. Motivated cognitive processing and attitude change. *Personality and Social Psychology Bulletin*, *24*, 715-726.

Mattijssen, E. J. A. T., Kerkhoff, W., Berger, C. E. H., Dror, I. E., & Stoel, R. D. (2016). Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination. *Science and Justice*, *56*, 113-122.

Merlino, M. (2015). Validity, reliability, accuracy, and bias in forensic signature identification. [Report to the United States Department of Justice].

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

Morris, R. (2000). *Forensic handwriting identification: Fundamental concepts and principles*. San Diego, CA: Academic Press.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, *30*, 395-406.

Nakhaeizadeh, S., Dror, I., & Morgan, R. (2014) Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. *Science & Justice*, *54*, 208-214.

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.

National Commission on Forensic Science. (2015). Ensuring that forensic analysis is based upon task-relevant information. *National Institute of Standards and Technology*.

National Commission on Forensic Science. (2017). Reflecting back – Looking toward the future. *National Institute of Standards and Technology*.

National Institute of Standards and Technology (2018). *The Organization of Scientific Area Committees for Forensic Science*. Retrieved from https://www.nist.gov/topics/forensic-science/organization-scientific-area-committees-osac

National Registry of Exonerations. (2018). *Exonerations in the United States*. [online] Retrieved from http://www.law.umich.edu/special/exoneration/Pages/Exonerations-in-the-United-States-Map.aspx

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: The National Academies Press.

Newman-Toker, D. E., & Pronovost, P. J. (2009). Diagnostic errors—the next frontier for patient safety. *JAMA, 301*, 1060-1062.

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

Northcraft, G. R., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39,* 84-97.

Office of the Inspector General. (2006). *A review of the FBI's handling of the Brandon Mayfield case*. US Department of Justice, Office of the Inspector General, Oversight and Review Division.

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

Osborn, A. S. (1929). *Questioned documents* (2nd ed.). Albany, NY: Boyd Printing Co.

Osborne, N. K. P., & Taylor, M. C. (2018). Contextual information management: An example of independent-checking in the review of laboratory-based bloodstain pattern analysis. *Science & Justice*, *58*, 226-231.

Osborne, N. K. P., Taylor, M. C., Healey, M., & Zajac, R. (2016a). Bloodstain pattern classification: Accuracy, effect of contextual information and the role of analyst characteristics. *Science and Justice*, *56*, 123-128.

Osborne, N. K. P., Taylor, M. C., & Zajac, R. (2016b). Exploring the role of contextual information in bloodstain pattern analysis: A qualitative approach. *Forensic Science International*, *260*, 1-8.

Osborne, N. K. P., & Zajac, R. (2015). An imperfect match? Crime-related context influences fingerprint decisions. *Applied Cognitive Psychology*, *30*, 126-134.

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 279-307). San Francisco, CA: Freeman.

Park, R. C. (2008). Signature identification in the light of science and expectation. *Hastings Law Journal*, *59*, 1101-1157.

Peschel, O., Kunz, S. N., Rothschild, M. A., & Mützel, E. (2011). Bloodstain pattern analysis. *Forensic Science, Medicine, and Pathology*, *7*, 257-270.

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.

Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology*, *69*, 309-315.

Podlas, K. (2005). The CSI effect: Exposing the media myth. *Fordham Intellectual Property, Media & Entertainment Law Journal*, *16*, 429-465.

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods* [Report]. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Pronin, E. (2006). Perception and misperception of bias in human judgment. *Trends in Cognitive Science*, *11*, 37-43.

Pronin, E., Yin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369-381.

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

Reingold, E. M., Charness, N., Pomplum, M., & Stampe, D. M. (2001). Visual span in expert chess players. *Psychological Science*, *12*, 48-55.

Risinger, D. & Loop, J. (2002). Three card monte, monty hall, *modus operandi* and "offender profiling": Some lessons of modern cognitive science for the law of evidence. *Cardozo Law Review*, *24*, 193-285.

Risinger, D., & Saks, M. (1996). Science and nonscience in the courts: Daubert meets handwriting identification. *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.

Robbers, M. L. (2006). Blinded by science: The social construction of reality in forensic television shows and its effect on criminal jury trials. *Criminal Justice Policy Review*, *19*, 84-102.

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.

Rumelhart, D. E. (2017). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (pp. 33-58). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Rumelhart, D. E., & Ortony, A. (1976). The representation of knowledge in memory. In R.C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99-135). Hillsdale, N.J.: Erlbaum.

Russo, F. D., Pitzalis, S., & Spinelli, D. (2003). Fixation stability and saccadic latency in elite shooters. *Vision Research*, *43*, 1837-1845.

Saber Tehrani, A. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. (2013). 25-year summary of US malpractice claims for diagnostic errors 1986-2010: An analysis from the National Practitioner Data Bank. *The BMJ*, *22*, 672-680.

Saks, M. J. (2009, October 30). Concerning L.J. Hall, E. Player, "Will the introduction of an emotional context affect fingerprint analysis and decision-making?" [Letter to the editor]. *Forensic Science International*, *191*, 19.

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

Sanna, L. J., Schwartz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology*, *28*, 497-502.

Schultz-Hardt, S., Frey, D., Luthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology*, *78*, 655-669.

Schwartz, N., & Vaughan, L. A. (2002). The availability heuristic revisited: Ease of recall and content of recall as distinct sources of information. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 103-121). Cambridge, UK: Cambridge University Press.

Schweitzer, N. J., & Saks, M. J. (2007). The CSI effect: Popular fiction about forensic science affects the public's expectations about real forensic science. *Jurimetrics*, *47*, 357-364.

Scientific Working Group for Forensic Document Examination (SWGDOC). (2013). Guidelines for forensic document examination. *Forensic Science Communications*, *2*. Retrieved from https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/april2000/swgdoc1.htm/swgdoc2.htm#Working Guideline for the Examination of Handwritten Items

Scientific Working Group for Forensic Document Examination (SWGDOC). (2018). Published Standards. Retrieved from http://www.swgdoc.org/index.php/standards/published-standards

Searston, R. A., & Tangen, J. M. (2017). Expertise with unfamiliar objects is flexible to changes in task but not changes in class. *Plos One*, *12*, e0178403.

Searston, R., Tangen, J., & Eva, K. W. (2015). Putting bias into context: The role of familiarity in identification. *Law and Human Behavior*, *40*, 50-64.

Shelton, D. E., Kim, Y. S., & Barak, G. (2006). A study of juror expectations and demands concerning scientific evidence: Does the "CSI effect" exist? *Vanderbilt Journal of Entertainment & Technology Law*, *9*, 331-368.

Sita, J., Found, B., & Rogers, D. (2002). Forensic handwriting examiners' expertise for signature comparison. *Journal of Forensic Sciences*, *47*, 1117-1124.

Singh, H., Meyer, A. N. D., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: Estimations from three large observational studies involving US adult populations. *BMJ Quality & Safety*, *23*, 727-731.

Smalarz, L., Madon, S., Yang, Y., Guyll, M., Buck, S. (2016). The perfect match: Do criminal stereotypes bias forensic evidence analysis? *Law and Human Behavior*, *40,* 420-429.

Snyder, M. (1981). On the self-perpetuating nature of social stereotypes. In D. L. Hamilton (Ed.). *Cognitive processes in stereotyping and intergroup behavior* (pp. 183-212). Psychology press.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*, 1202-1212.

Snyder, M., & Uranowitz, S. W. (1978). Reconstructing the past: Some cognitive consequences of person perception. *Journal of Personality and Social Psychology*, *36,* 941-950.

Srihari, S., Cha, S., Arora H., & Lee, S. (2002). Individuality of handwriting. *Journal of Forensic Sciences*, *47*, 856-872. doi: 10.1109/ICDAR.2001.953764

Stacey, R. (2004). A report on the erroneous fingerprint individualization in the Madrid train bombing case. *Journal of Forensic Identification*, *54,* pp. 706-718.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137-149.

State v. Barnhouse, 18D02-9204-CF-000046 (Ind. 1994).

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*, 523-540.

Steinback, R. (2007). The fight for post-conviction DNA testing is not yet over: An analysis of the eight remaining holdout states and suggestions for strategies to bring vital relief to the wrongfully convicted. *Journal of Criminal Law & Criminology*, *98*, 329-361.

Steinman, M. A., Shlipak, M. G., & McPhee, S. J. (2001). Of principles and pens: Attitudes and practices of medicine housestaff toward pharmaceutical industry promotions. *The American Journal of Medicine*, *110*, 551-557.

Stevenage, S. V., & Bennett, A. (2017). A biased opinion: Demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results. *Forensic Science International*, *276*, 93-106.

Stoel, R., Berger, C., Kerkhoff, W., Mattijssen, E. J. A. T., & Dror, I. (2014). Minimizing contextual bias in forensic casework. In K. J. Strom & M. K. Hickman (Eds.), *Forensic science and the administration of justice: Critical issues and directions* (pp. 67-86). SAGE Publications, Inc.

Stuart, J. O. R., Windschitl, P. D., Smith, A. R., & Scherer, A. M. (2017). Behaving optimistically: How the (un)desirability of an outcome can bias people's preparation for it. *Journal of Behavioral Decision Making*, *30*, 54-69.

Sulner, A. (2014). Handwriting: Cognitive bias. *Wiley Encyclopaedia of Forensic Science*.

Sun, H-M., Simon-Drack, S. L., Gordon, R. D., & Teder, W. A. (2011). Contextual influences on rapid object categorization in natural scenes. *Brain Research*, *1398*, 40-54.

Tangen, J. A., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*, 995-997.

Thomas, E. J., Studdert, D. M., Burstin, H. R., Orav, J. E., Zeena, T., Williams, E., Howard, M. K., Weiler, P. C., & Brennan, T. A. (2000). Incidence and types of adverse events and negligent care in Utah and Colorado. *Medical Care*, *38*, 261-271.

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: Experts can do a lot with a little. *Plos One,* e114759.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in fingerprint identification. *Journal of Forensic Sciences*, *58,* 1519-1530.

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.

Thompson, W. C. (2011). What role should investigative facts play in the evaluation of scientific evidence? *Australian Journal of Forensic Science*, *43*, 123-134.

Thompson, W. C., & Cole, S. A. (2005). Lessons from the Brandon Mayfield case. *The Champion*, *29*, 32-34.

Thornton, J. (2010). A rejection of "working blind" as a cure for contextual bias [Letter to the editor]. *Journal of Forensic Sciences, 55*, 1663.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*, 47-58.

Towler, A., White, D., Ballantyne, K., Searston, R. A., Martire, K. A., & Kemp, R. I. (2018). Are forensic scientists experts? *Journal of Applied Research in Memory and Cognition*, *7*, 199-208.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*, 109-123.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, *7*, 1-12.

Wadsworth, B. J. (2004). *Piaget's theory of cognitive and affective development: Foundations of constructivism*. Longman Publishing.

Wänke, M. (2013). Almost everything you always wanted to know about ease-of-retrieval effects. In C. Unkelback, & R. Greifeneder (Eds.). *The experience of thinking: How the fluency of mental processes influences cognition and behaviour* (pp. 151-170). Psychology Press.

Wason, P. C. (1960). On the failure to eliminate hypothesis in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*, 129-140.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*, 1049-1062.

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 and Cognition*, *2*, 53-55.

Whitman, G., & Koppl, R. (2010). Rational bias in forensic science. *Law, Probability, & Risk*, *9*, 69-90.

Wickham, P. A. (2003). The representativeness heuristic in judgments involving entrepreneurial success and failure. *Management Decision*, *41*, 156-167.

Windschitl, P. D., Scherer, A. M., Smith, A. R., & Rose, J. P. (2013). Why so confident? The effects of outcome desirability on selective exposure and likelihood judgments. *Organizational Behavior and Human Decision Processes*, *120*, 73-86.

Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, *44*, 68-82.

Zadny, J., & Gerard, H. B. (1974). Attributed intentions and informational selectivity. *Journal of Experimental Social Psychology*, *10*, 34-52.

Zajac, R., Barrett, E., Osborne, N., Hegermann, R., & Kouwenhoven, M. K. J. (2018). The fine print: Disentangling the effects of crime context and emotion on fingerprint decisions. [Manuscript under revision].

Zegers, M., de Bruijne, M. C., Wagner, C., Hoonout, L. H. F., Waaijman, R., Smits, M., Hout, F. A. G., Zwaan, L., Christiaans-Dingelhoff, I., Timmermans, D. R. M., Groenewegen, P. P., & van der Wal, G. (2009). Adverse events and potentially preventable deaths in Dutch hospitals: Results of a retrospective patient record review study. *BMJ Quality & Safety*, *18*, 297-302.

Appendix A: Interview protocol for Study 3.

Background

Before we start, can you please tell us how long you have been working as a forensic document examiner and describe any training you have done for this type of work.
What type of agency do you work within (e.g., private, police, independent, academic)?
What is your educational background? What qualifications do you require to become a QDE?
Research background?
Proficiency testing? Feasibility of in-house testing administered by context manager as a "real" case – blinding examiner to the fact that it is a test

QDE Methods

Who requires your expertise? How are you contacted? How often?

Breaking down the stages of QDE

In your own words, describe the role of the forensic handwriting examiner, and the nature of your expertise?
Describe the various roles within QDE that you may have (e.g., lab manager, bench analyst, reviewer)
Are you an expert in any other forensic disciplines?
Can you please give a brief description of the document examination process? Of the handwriting examination process (if different)? Signature examination (if different)?
What is an adequate number of exemplars that you need to compare to the questioned document?
What instruments/tools do you need to begin an examination?
From what we know, there appear to be four main types of cases in which you might be required to do a signature comparison/handwriting examination.

- ***Alleged forgery*** *(in which you might compare the questioned writing/signature to that of a suspect and/or purported writer, anyone else?)*
- ***Alleged disguise*** *(e.g., to avoid culpability)*
- ***Who done it?*** *(e.g., threatening letter/ransom (open set) or incriminating statement (e.g., from prisoner, closed set))*
- ***Inner comparison*** *– were two questioned documents written by the same person? Where there is no "known" writer*
- *Can you give us some examples of cases like those described above where your expertise were required?*
- *Are there any other case types that you might be involved in? Can you give us some examples?*
- *Would you ever be asked, for example, to comment on the* ***mental/physical state of the writer****?*
- *What about the* ***nature of the writing instruments/paper****?*

From what we know, the main opinion that you will offer is about source (i.e., whether a suspect did or did not do the questioned writings). Is that correct?

- *Genuine, not genuine (don't talk about in terms of forgery)*

Can you please describe the opinion language and levels of those opinions that you use (e.g., likelihood ratios, probabilities, verbal scale)?
[Discuss **genuine**, **not genuine**, **same source but disguised to look different** if participant doesn't talk about these.]
What is your role in informing police and lawyers as to the QDE conclusions? Would you ever be asked to help develop theories of the case for the police? For example, how or why an action might have occurred?
Has your expertise helped the broader context of the case? How was this done? Can you help to determine possible motives? Under what circumstances does this happen? Would this be done before or after or within your report?

General assessment of information in QDE

Prior to examination of the documents, what information do you typically need in order determine whether or not you will accept a case?
In what circumstances would you need information beyond the handwriting sample? What would this information be?
What is the minimum amount of information needed to conduct an examination? Does this differ depending on the case? In what way?
What information do you typically get that you don't *need? Who gives you this information? Why?*
Do you ever receive information that you think could have a biasing effect on your examination? How do you deal with this information?
In some cases (for example a handwriting sample that contains incriminating evidence) there may be contextual information contained within the sample. How do you deal with this information? Is it important for your analysis? Do you take any steps to avoid this type of information?

Information relevance at each stage of Handwriting Examination Process

Are you aware of the NIST Handwriting Examination Process Map? Do you use this as a methodology?
If not, do you have another standard that you follow? What is it?
We are now going to talk through the NIST process map, and will ask what information you need at each stage, if any.

Questioned writing pre-analysis.

- *(0) What information do you need to review and verify the documents sent for examination? Does this vary depending on the case?*
- *(110) How do you determine if the Q document has sufficient clarity and detail? What information do you need at this stage other than the Q document?*
- *(160) How would you determine if consultation/research would assist in your examination?*
- *(120) What do you do if you are not familiar with the character set or signature style used?*
- *(130) How do you determine if there is enough Q material to compare? What information do you need at this stage other than the Q documents?*

Known writing pre-analysis.

- *(220) How do you determine if the known sample contains sufficient clarity and detail? What information do you need other than the K sample(s)?*

- *(230) How do you determine if there is enough comparable K writing?*

Does the same analyst who determines the sufficiency of Q and K for analysis also perform the analysis? Why, why not?

Analysis of Questioned writing.

- *(310) When you observe the characteristics in the Q document, what other information (if any) do you need to do this?*
- *(320) How do you assess the level of complexity? What other information do you need?*
- *(330) What information do you need to consider the type of document that you are looking at?*
- *(340) What information do you need to consider the writing instruments?*
- *(350) How do you determine if grouping the Q samples by observed similarities would be helpful?*
- *(370) How do you formulate initial possible explanations for observations of Q document?*

Analysis of Known writing.

- (*400) When you observe the characteristics in the K writing, what other information (if any) do you need to do this? Will you have the Q sample near you at this point?*
- *(410) How do you assess the level of complexity? What other information do you need?*
- *(430) What information do you need to consider the type of document that you are looking at?*
- *(420) What information do you need to consider the writing instruments?*
- *(440) What information do you need to consider whether K is contemporaneous (made at the same time) with Q?*
- *(450) How do you assess K samples for multiple writers (inter-se comparison). What information do you need to do this?*
- *(460) How do you formulate initial possible explanations for observations of K document?*
- *(350) How do you determine if grouping the K samples by observed similarities would be helpful?*
- *(510) How do you assess the range of variation of the K writer?*
- *(520) How do you determine if the K samples are reliable and suitable for comparison?*

Comparison of Q and K writing samples.

Do you compare the Q and K at the same time? If you haven't been involved in the pre-analysis, will you see the K writing before/at the same time as the Q writing?
Do you see any issues/advantages with seeing the Q and K writing at the same time?

- *(530) How do you determine which characteristics are of interest? What information other than the writing samples do you need to determine this?*

- *(540) How do you classify distinguishing features as dissimilarities and similarities among corresponding characteristics? What information do you need to do this?*
- *(550) What information do you need to note absent characteristics?*
- *(560) How do you determine if dissimilarities are due to natural variation? What information other than the Q and K samples do you need to determine this?*

Evaluation.

- *(600) What information do you need in order to determine if there are limitations present in the quantity, quality, and comparability of the samples?*
- *How do you determine what are* ***significant*** *differences?*
- *How do you determine what are* ***sufficiently significant*** *differences?*
- *How do you determine what are* ***significant distinguishing characteristics****?*
- *How do you determine what are* ***significant*** *similarities?*
- *How do you determine what are* ***sufficiently significant*** *similarities?*
- *Do you use case information to help you separate the difference between* ***significant, sufficiently significant****, and* ***significant distinguishing characteristics****?*
- *How do you determine if the combination of significant distinguishing characteristics are substantial enough to make an identification?*
- *Inconclusive vs. no conclusion – do you differentiate between these in your report? How? If not, why?*
- *Opinion on using Likelihood Ratio vs. categorical identification decisions. There is a strong push in all forensic disciplines to move away from categorical-identification classifications. What is your opinion of this?*

Review

Are your conclusions reviewed? By who? (If no, why not?)
Can you please describe the review process?
Does the reviewer know your conclusions before the review is conducted?

Report/Court

Does the format of your report differ depending on who has requested your expertise? In what way?
Does the format of your report differ depending on the type of case? In what way?
Does your report inform the reader of all the information other than the writing samples that you had access to?
What is your role in Court? How often is your report used in Court proceedings? Is this usually in person or do you have statements read by a Clerk?
How is your report used if not in Court? By whom?

Specific Types of information

Now we are going to list specific types of information that you might encounter on any given case. Could you please discuss whether (and why) this information is relevant, irrelevant, whether the relevance depends on the case (and how), and whether it is possible to avoid this information?

- Who is requesting your services
- The requester’s theory about the writing samples

- Alternative theories of the writing sample (i.e., police have requested your services believing X happened, but defence are arguing that Y happened)
- The suspect's name and details
- The name and details of any other parties
- The suspect's criminal history (e.g., previously committed forgery)
- The case name or number
- The content of the writing (e.g., in a random note/threatening letter)
- How your conclusion will be used by the client
- How urgent the case is; i.e., how quickly your conclusion is required
- ? [list more here]

Are there any other types of information that you can think of that would be valuable for you to consider during your examination?

Context management

Do you employ context management procedures at present? Why/why not? What barriers are/were there for you to do so?

Do you think context management is the most appropriate way of dealing with unconscious bias? Why/why not?

To finish

Would you be interested in participating in an experimental study further examining some of the issues identified?

National Commission definition of *task-relevance*. [This is here for reference if the definition comes up in discussion]

Information is *task-relevant* for analytic tasks if it is necessary for drawing conclusions:

1) About the propositions in question
2) From the physical evidence that has been designated for examination
3) Through the correct application of an accepted analytic method by a competent analyst.

Information is *task-irrelevant* if:

1) It is not necessary for drawing conclusions about the propositions in question.
2) It assists only in drawing conclusions from something other than the physical evidence designated for examination; or
3) If it assists only in drawing conclusions by some means other than an appropriate analytic method.