The Number of Unanimous Witnesses who Identify a Suspect from a Lineup Influences Mock Jurors

by

Robbie J. Taylor

A thesis submitted to Victoria University of Wellington in fulfilment of the requirements for the degree of Doctor of Philosophy

Victoria University of Wellington

2021

Abstract

During a criminal proceeding, jurors need to weigh up the presented evidence and determine a verdict. Research has shown that witness identification evidence is compelling to jurors, despite the fact that it can be unreliable. How reliable are the combined lineup decisions gathered from multiple witnesses? Generally, the more witnesses who identify the same person from a lineup, the more likely that person is guilty. But recent theoretical evidence suggests that a greater number of witnesses identifying the same person from a biased lineup can indicate that person is actually *less* likely to be guilty than if there were a smaller number of witnesses identifying that person (Gunn et al., 2016). As the number of agreeing witnesses increases, the more likely that agreement is caused by the lineup bias, rather than consistent witness memories of the crime. In this thesis, I examined how unanimity and lineup bias influenced jurors' perceptions of guilt. Subjects who saw a biased lineup gave lower ratings of guilt compared to subjects that were shown a lineup that had no obvious bias. In addition, warning subjects that a lineup was biased led them to give lower guilt ratings than subjects who did not receive a warning. Subjects who were told there were two witnesses who identified the police suspect gave higher guilt ratings than subjects who were told there was one witness who identified the police suspect, but only when the lineup was clearly not biased. Subjects' guilt ratings were not significantly greater in conditions with more than two unanimous witnesses identifying the police suspect. It seems subjects had a limit of certainty based on changes in witness numbers alone. We also found that the way in which witness numbers were presented to subjects influenced guilt ratings. When we presented witnesses coming forward in different

groups and on different days, subjects shifted their guilt ratings upwards. When the number of witnesses decreased during the experiment, subjects did not decrease their guilt ratings to the same extent as those subjects in conditions in which the number of witnesses increased by the same magnitude. This finding is consistent with the literature on *confirmation bias* and the *story model* of juror decision-making—subjects likely formed an initial belief that the identified suspect was guilty and subsequent evidence was evaluated against that belief (Nickerson, 1988; Pennington & Hastie, 1993). The finding that presenting witnesses coming forward in separate groups increased subjects' guilt ratings adds to the literature showing that jurors are influenced by irrelevant information presented to them during a proceeding. This research also demonstrates that future research should examine strength of evidence manipulations over multiple levels—rather than as dichotomous "strong" and "weak" extremes.

Acknowledgements

Thank you to Victoria University of Wellington for funding through the Victoria Doctoral Scholarship, Doctoral Submission Scholarship, and Strategic Faculty Grant. Thanks to Maryanne Garry for teaching me science is hard and for demanding the highest standards. Also thanks to those in the Garry lab for their advice and support in the early stages of my PhD. Thanks to Matt Crawford for your guidance and agreeing to supervise me at the later stages of my PhD. Thanks to Gina Grimshaw for your advice and mentorship. Thanks to Melissa Colloff and Kim Wade for their excellent lineup materials. Thanks to my family for support and financial assistance over the years. Finally, thanks to Diana for your patience over the years.

Contents

List of Tables and Figures

Chapter 1

The Case of Adolf Beck

On 16 December 1895, as Adolf Beck exited his London apartment, a woman named Ottilie Meissonier approached him and said, “I know you” (Cathcart, 2004). Meissonier claimed Beck was the man she had met three weeks earlier who had introduced himself as Lord Wilton—a wealthy man who had recently inherited an estate. According to Meissonier, Lord Wilton was charming, and had offered her the job of housekeeper of his estate and to be his mistress in exchange for expensive clothing, jewellery, foreign travel, and accommodation. Meissonier also claimed Lord Wilton had given her a cheque to buy clothing, but when she attempted to use the cheque, it was invalid. Meissonier had given Wilton some of her rings and watches after he promised that he would use them to buy new jewellery in her size. But she never saw the man again until, unexpectedly, the morning of 16 December. Meissonier was certain that Beck was the con artist going by the name of Lord Wilton. Beck was confused by Meissonier’s accusations and swiftly walked away from her. Meissonier was insistent and followed Beck, weaving through carts and carriages of the London streets, shouting after him. After this continued for a time, both Meissonier and Beck ran to a police officer in the street. Meissonier told the officer she was confident that Beck was the con artist who stole her jewellery. The police took both Meissonier and Beck to the station. As it turned out, the police had received 22 similar complaints in the previous two years about a con artist going by the name of Lord Wilton with the same modus operandi.

To verify that Beck was indeed the con artist, the police constructed multiple physical lineups by finding 10-15 men from the streets of London who they thought

resembled Beck and asked them to stand in a line next to Beck (Cathcart, 2004; Dicks, 2007). The police then asked the 22 victims to view the lineups and decide if Lord Wilton was one of the members of the lineup. All 22 women identified Beck from the multiple different lineups. Beck was found guilty of fraud because, according to the judge, "the evidence of identity has been absolutely overwhelming." Police also successfully argued that Beck was responsible for a series of crimes that occurred in London 18 years earlier with the same modus operandi (i.e., a con artist by the name of Lord Wilton taking jewellery from women). A man named John Smith had been convicted for these earlier crimes, but went missing after the conviction. Police successfully argued that Beck and Smith were, in fact, the same man: Lord Wilton. Beck was also charged with these earlier crimes.

In 1901, after five years in prison, Beck was paroled for good behaviour at the age of 60. Three years later, a woman named Pauline Scott complained to police that a man named Lord Wilton had stolen a ring, watch, and a sovereign. The police recognised the story and connected it to Beck (Collins, Walpole, & Edge, 1904, cited in Clark, Moreland, & Rush, 2015; Cathcart, 2004). Scott later identified Beck from a lineup and the publicity of the crime brought forward four other women who were also victims of similar crimes, and who also identified Beck from a lineup. Despite Beck proclaiming his innocence, the weight of multiple witnesses identifying Beck was enough to once again send him back to prison.

Ten days after Beck's second conviction—and while he was in prison—a similar crime occurred in London: A man going by the name of Lord Wilton took jewellery from two sisters. Unlike the previous crimes, the sisters were suspicious of Wilton and asked

their landlord to follow him. The landlord found Lord Wilton at a pawnshop attempting to sell the sisters' jewellery and he contacted the police, leading to Wilton's arrest. The man was identified as John Smith and he confessed to the earlier frauds. Beck was subsequently released and awarded £5,000 as compensation for his wrongful convictions.

In total, over his two convictions, 27 women independently identified Beck from multiple different lineups. How is it possible that 27 independent eyewitnesses could be wrong? Intuitively, one would expect that as the number of witnesses identifying the same person from a lineup increases, the probability that the identified person is actually guilty would also increase. But in this case, none of these 27 women realised that Beck was, in fact, innocent. Indeed, there have been many cases in which witnesses have wrongly remembered a crime or the people involved in a crime. In the United States, approximately 70% of wrongful convictions which were later exonerated with DNA evidence involved mistaken witness identification (Innocence Project, 2020; West & Meterko, 2015, see www.innocenceproject.org for examples). In attempting to understand *how* errors in memory for events and faces happen, researchers have examined the specific circumstances of these cases, and then replicated the sorts of witness errors in simulated circumstances in the laboratory. The evidence is unequivocal—eyewitness memory is unreliable evidence (Cutler, Penrod, & Martens, 1987; Koriat, Goldsmith, & Pansky, 2000; Loftus, 1981; Schacter, 1999). This unreliability is often due to normal memory processes. That is, details stored in memory are not complete photographic records of the past. Instead, recalling an event involves reconstructing that event, which relies on general knowledge, heuristics, and event-relevant details to piece together a coherent narrative of what happened (Bartlett, 1932; Kolodner, 1983; Greene & Loftus, 1984; Loftus, 2005;

Rumelhart & Ortony, 1976; Rumelhart, 2017). General knowledge structures, referred to as *schema,* are used to fill in details of specific memories when those details are missing due to a failure to encode or retrieve them (Bartlett, 1932). Errors can slip into how the event is *remembered* at any point during this reconstructive process and errors in the recall at one point in time can become part of the memory itself because people often fail to identify and differentiate the source of introduced details from details of the original event (Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981; Lindsay, 1994).

Although this research might be able to explain how the memories of the 27 women in Beck's case might be inaccurate, the research does not necessarily explain why *all 27 women* identified Beck from the lineup. If these women were truly independent witnesses one would expect their constructive memory processes to be varied. That is, one might expect the women from varied backgrounds to draw on slightly different schemas, heuristics, and autobiographical memories to "fill in" missing or forgotten parts of their memories of the crime (Hyman & Billings, 1998). The fact that all of them recreated an event in which Beck was without doubt the con artist seems inexplicable. Additionally, one might expect that after these sometimes lengthy and intimate interactions with Wilton, some of the women might have correctly realised Beck was innocent. That is, this was not a case of a quick crime that happened with no meaningful interaction with the perpetrator—there was ample time and (presumably) motivation to process the event carefully. Indeed, the length of exposure to an unfamiliar face increases subsequent identification accuracy (Bornstein, Deffenbacher, Penrod, 2012; Reynolds & Pezdek, 1992). It is therefore puzzling that so many women were in agreement over the identity of Beck.

The unanimity in the women's identification decisions suggests that they could have been influenced systematically. For example, all of the witnesses could have been influenced by the way the lineups themselves were constructed. In Beck's case, many of the witnesses reported that they thought Beck was the only person who closely resembled the perpetrator. That is, Beck stood out as being the only plausible choice (Cathcart, 2004; Clark, Moreland, & Rush, 2015). If this were true, then perhaps the victims simply chose the person (Beck) who most resembled the true perpetrator (Smith), leading to unanimity amongst the 27 women. Indeed, evidence from studies involving mock witnesses evaluating evidence from simulated crimes demonstrates that witnesses are likely to identify individuals from lineups who *look most like* the perpetrator—regardless of whether or not the actual perpetrator is present (Steblay & Dysart, 2001; Wells, 1984; Wells et al., 1998). This strategy of identifying the person who most closely resembles the actual perpetrator is easier for witnesses if the other people in the lineup do not closely resemble the perpetrator.

In this thesis, we[1] attempted to simulate a situation similar to Beck's case in which witnesses unanimously identified the same person—who clearly stood out—from a lineup. The purpose of this thesis was to understand the relationship between people's beliefs about guilt when the number of unanimous witnesses identifying the same person from a

[1] I have used the word "we" here and other instances in this thesis because although this thesis is my own work, the research was conducted with the advice from members of my research lab and my supervisors. In contrast, I have used the word "I" when I am referring to my writing, actions, or my personal understanding of the research literature.

lineup varied. To limit the scope of the current research programme, we were not aiming to replicate the exact circumstances of a real trial-by-jury. That is, participants in these studies were asked to make guilt likelihood ratings as if they were a juror considering the evidence, but none were part of a *mock jury* procedure which would have involved the consideration of additional evidence, arguments, and group discussion. Further research will be required to establish the extent to which the findings from the experiments presented in this thesis apply to real trials.

In this chapter, I will first explain the historical origins of the *trial-by-jury* and literature on how jurors evaluate evidence presented during a court proceeding. I will then focus on the specific evidence used in Beck's case—*lineups*. I will explain the origins and purpose of lineups and provide an overview of the psychological literature on lineups. Then, I will describe the research on biased lineups, including the definition of biased lineups, how one might measure lineup bias, and how jurors interpret biased lineups. The final pertinent aspect of Beck's case was the presence of multiple witnesses. I will therefore describe the research examining multiple witnesses, and then explain how jurors might interpret multiple witness evidence from a biased lineup. Finally, I will review the literature on the effects of judicial warnings on jurors' evaluation of evidence.

History of the Trial-by-Jury

The trial-by-jury is a widely adopted legal procedure in many countries—often regarded as a hallmark of justice and civil participation (Brooks, 2004; Hamilton, 1961). Some historians have attributed the development of the trial-by-jury to Anglo-Saxon origins, or as a result of the Norman conquest, which then further developed in England (Forsyth, 1875). However, despite the strong ties to the English judicial system, elements

of the modern trial-by-jury process were also documented earlier in Scandinavian, Roman, and German legal systems (Forsyth, 1875; Moschzisker, 1921). Therefore, the trial-by-jury process cannot be attributed to one time, place, or person—it naturally evolved over centuries.

In the early Middles Ages, before trial-by-juries were widely adopted, *trial-by-ordeal* was a popular legal process to resolve disputes (White, 1961). Instead of hearing and weighing up evidence to convince a judge of the truth, the trial-by-ordeal involved a dangerous and often life-threatening task, that if performed by the defendant indicated an exoneration by God. One common task involved blindfolded people walking down a path with scattered pieces of red-hot iron. If the person could reach the other side without being burned by the iron, then the accused would be considered innocent. Over time, however, the trial-by-ordeal evolved and some of the added elements began to resemble features of the trial-by-jury.

One example of the transition between trial-by-ordeal and trial-by-jury is evident in the English justice system. King Henry II introduced the *assize* in 1166 as a court trial for primarily land disputes. The trial involved twelve "of the more lawful men" who were summoned to determine the rightful owner of the land. These jurors, and most jurors in the Middle Ages relied on local knowledge, character judgments, and out-of-court interactions with defendants to determine the facts of the case (Klerman, 2018). In other words, the jurors acted as a screening process in deciding whether or not the accusations were credible. If the accusation did not seem credible, the defendant was exonerated. If the accusation was credible, the defendant underwent trial-by-ordeal—in which God was to determine the defendant's fate.

By 1215 the trial-by-ordeal came to an end due to a lack of support from the church (Kerr, Forsyth, & Plyley, 1992) and trial-by-jury was chosen as its replacement. Because jurors already gave verdicts as part of a screening process in trial-by-ordeals, the main change was to make these verdicts final instead of subjecting the defendant to ordeal. That is, the trial-by-jury represented a shift in the process from judgment by God to judgment by peers. Juries also became less self-informing throughout the fourteenth and fifteenth centuries—although it is not entirely clear what caused this change. This decline in self-informing jurors (i.e., jurors who use their own local knowledge to decide guilt or innocence) led to the development of prosecutors, who presented evidence to the jury (Green, 1988; Klerman, 2018). The move away from self-informing jurors also led to an increase in witness testimony in court.

Over time, the English trial-by-jury evolved into the modern jury system. For instance, in 1352, the defendant could challenge who was on their jury; in 1367, it was decided that the twelve jurors had to reach a unanimous rather than majority decision (Anand, 2005). Furthermore, cases between fifteenth and eighteenth centuries were lawyer-free, referred to as *altercation trials*, which were largely unstructured hearings of both side's accounts (Langbein, 2003). In these altercation trials, defendants were not assumed to be innocent until proven guilty beyond reasonable doubt. Instead, if the defendant was innocent, then the defendant was expected to demonstrate his or her innocence to the jury. It was not until late eighteenth century that the beyond reasonable doubt burden of proof was used—which came with the introduction of the *adversarial trial*. With the change to the adversarial trial was the reduction, and eventual elimination, of jurors questioning witnesses and requesting further evidence in court (Klerman, 2018).

From these historical origins, the modern trial-by-jury still remains today in many legal jurisdictions. Why is the trial-by-jury the dominant approach to criminal justice today? Perhaps the procedure is still used, in part, because of its long history and tradition in society. The jury system does promote civic participation and represents current values of society (Crosbie, 2020). It is certainly true, however, that the jury system does not remain because it is a flawless method for determining the truth. Indeed, using "everyday people" to determine guilt or innocence and administer the law can have major disadvantages. Specifically, jurors are simply not trained in the use of evidence in determining truth, they often have little knowledge of the legal system, are often swayed by irrelevant information, and susceptible to myriad biases and errors (Desmarais & Read, 2011; MacCoun & Kerr, 1988; Mazzella & Feingold, 1994; Reifman, Gusick, & Ellsworth, 1992; Woods, 2018).

Models of Juror Decision-Making

Given the general lack of training in the consideration of evidence and judicial practice, how *do* jurors evaluate defendant culpability during a trial? To help answer that question, it is useful to break down the two decision-making components of a trial-by-jury. The first component is the individual juror pre-deliberation process, during which jurors hear and weigh-up evidence to come to an initial decision of guilt. The second component is the jury deliberation process, in which jurors discuss together the evidence as a jury to reach a verdict. Most research in the literature has examined one or the other of these processes; however, there have been some attempts to integrate the separate literatures on both processes (Levett, Danielsen, Kovera, & Cutler, 2005; Reardon, 2008; Schmersal, 2011). In this thesis, we examined the first process, individual pre-deliberation, to

understand how people evaluate evidence to determine the probability of guilt and how that probability is updated during the evidence presentation portion of a trial scenario. To understand how jurors determine the probably of guilt, we first must understand how jurors generally evaluate evidence. In the juror-decision making literature, there are two broad classes of models that describe how jurors evaluate evidence: *mathematical models* and *explanation-based models*.

Mathematical models. Mathematical models describe juror decision-making as a result of predictable and logical calculations. Within these models, jurors are assumed to individually evaluate each piece of evidence presented, and then make a determination of guilt based on some mental calculation representing the combined strength of those pieces of evidence (Penrod & Hastie, 1979; Winter & Green, 2007). For example, the *Bayesian model* assumes that jurors start with an initial assessment of guilt (at the beginning of the trial), referred to an *a priori belief* (Devine, 2012; Gelfand & Solomon, 1973). This a priori belief would be based on, amongst other things, factors like the jurors' attitudes toward police and the criminal justice system, and judges' preliminary instructions (Hastie, 1993). This a priori belief is then updated during the trial with each new piece of evidence. Once all of the evidence has been presented, jurors compare their final determination of guilt, referred to as a *posterior belief*, to a cut-off point to determine their verdict. This cut-off point, or *criterion*, is the personal level of certainty required for a juror to decide on a guilty verdict (Devine, 2012). If jurors' final guilt ratings do not meet the threshold for a guilty verdict, then their initial verdict, before the deliberation process, will be *not guilty.* The cut-off threshold for each juror is likely to vary and might be influenced by factors such as judges' instructions.

Another class of mathematical models, similar to the Bayesian model, are *weighted linear models*—which describe juror decision-making as involving the use of algebraic equations (Devine, 2012; Ostrom, Werner, & Saks, 1978; Pennington & Hastie, 1981). According to these models, jurors consider both the weight and implication of each new piece of evidence. In this context, "weight" of evidence is determined by factors such as the reliability, credibility, and relevance of evidence (Hastie, 1993). The more weight placed on an individual piece of evidence, the greater impact that evidence will have on the jurors' final decisions. As in the Bayesian model, jurors generate a final guilt rating and then compare this rating to a subjective cut-off point. If the guilt rating is past this criterion, then a guilty verdict is made, otherwise, a not guilty verdict is made.

There have been many criticisms of these mathematical models (Hastie, 1993). The Bayesian model, for example, is criticised because it assumes that evidence is evaluated on a single measure (Pennington & Hastie, 1981). That is, evidence is evaluated to determine a posterior belief of guilt. However, in reality, it is likely that people evaluate evidence not only on how it influences the probability of guilt, but also on its reliability and trustworthiness. Another criticism of mathematical models is that jurors are unlikely to make one global judgment assessment of guilt after hearing the evidence (MacCoun, 1989). Instead it is likely that jurors make multiple decisions about a crime. For example, to what extent does the evidence show that the defendant was in the vicinity of the crime scene; to what extent does the evidence support the defendant's motive and intent to commit the crime? It is likely these are separate judgments that are determined by combining evidence. Finally, one of the most compelling criticisms of mathematical models is their simplicity. Many of these models do not account for situational or

contextual elements of the trial known to impact juror decision-making, such as the order in which evidence is presented, individual differences, and different types of cases (e.g., murder or theft) (Costabile, & Klein, 2005; Kerstholt, & Jackson, 1998; Pennington & Hastie, 1981).

Explanation-based models. To address some of the criticisms of mathematical models, researchers have turned away from attempting to describe juror decision-making as involving the mental computation of complex and rational equations to a focus on how jurors create a cognitive representation of the crime, using evidence and their knowledge of the world. One of the major differences between mathematical and explanation-based models is that unlike most mathematical models, explanation-based models do not assume jurors are passive evaluators of evidence, instead they assume that jurors use their own knowledge and assumptions to create a representation of the event.

The *story model* proposed by Pennington and Hastie (1981, 1986, 1988, 1993) states that jurors create a narrative of the crime based on the evidence presented during a trial. Jurors use the presented evidence about the case and combine that evidence with their own knowledge of similar cases and crimes and also general knowledge about how the world works. Because narratives are partly based on the jurors' idiosyncratic knowledge and personal biases, it is possible to create many different stories of what happened from the same evidence. The model assumes that jurors consider different plausible stories created from the evidence and then decide which one is the *most* plausible. Jurors decide on the most plausible option by considering each story's *coverage*, *coherence*, and *uniqueness* (Pennington & Hastie, 1991). Jurors then learn about the different verdict options during the judge's instructions to the jury. Finally, jurors use their constructed

narrative and evaluate it against the verdict options as described by the judge to decide which verdict best matches their story—referred to in the model as *goodness of fit*.

Importantly, the story model departs from the idea that jurors are rational and logical decision-makers. The model does not assume that jurors wait until hearing all of the evidence before creating their narratives. Instead, empirical evidence demonstrates that people arrive at a coherent story relatively early in the presentation of evidence. Once people have a preferred story, they tend to evaluate subsequent evidence in terms of whether or not it supports that story (Carlson & Russo, 2001; Russo, 2014; Russo, Medvec & Meloy, 1996). This process also involves confirmation bias as jurors tend to place more weight on new evidence that supports their preferred story than new evidence that is inconsistent with that story.

Another explanation-based model of juror decision-making is the *heuristic-systematic model* (Chaiken, 1980, 1987; Chen & Chaiken, 1999). The model is based on the idea that there are two ways people process information. The first way, involves the use of heuristics, or "rules of thumb" and the second involves careful evaluation and scrutiny. The heuristic-based processing, referred to as *Type I processing,* is quick and efficient. In contrast, the more deliberative processing, referred to as *Type II processing,* is more time-consuming and requires more cognitive effort (Evans, 2003, 2008; Kahneman & Frederick, 2002; Stanovich, 1999). Chen and Chaiken proposed that people evaluate evidence along a heuristic/systematic continuum. In the context of a court proceeding, systematic processing (Type II) is more desirable because it is assumed to be less susceptible to bias than Type I processing, but it also requires jurors to be motivated enough to deliberate and also to have the ability to interpret complex evidence (Chaiken & Ledgerwood, 2011). When evidence

becomes too difficult to evaluate systematically, or jurors lack the motivation to evaluate the evidence systematically, they are likely to resort to the more efficient heuristic-based processing (Type I). The heuristic-systematic model is similar to the *elaboration likelihood model* (ELM) of persuasion and attitude change (Petty & Cacioppo, 1986). The ELM is also a *dual-process model* with a systematic (or central) route involving effortful evaluation of evidence and a second heuristic-based (or peripheral*)* route involving simple inferences being drawn about the person, or source, of an argument, rather than the quality of the argument. Of course, heuristic processing is not necessarily bad—many times quick rules of thumb can be used to come to correct decisions. But heuristic processing can also involve jurors using irrelevant information, such as defendant attractiveness, and stereotypes to decide on a verdict (Kovera, McAuliff, & Hebert, 1999; Lieberman, 2002; Saks & Kidd, 1980).

Explanation-based models apply cognitive processes to the jury context, which have undoubtedly prompted much of the empirical research on how people actually evaluate evidence in real and simulated trials. However, it must be acknowledged, these models do not propose any novel cognitive processes, which is most obvious in the heuristic-systematic model because it is clearly based on Type I and II processing and the ELM model. It is also the case that the individual elements of the story model can be explained by basic cognitive processes and biases. Most notably is the process of confirmation bias: When jurors have established a narrative of what happened they tend to emphasise evidence that supports that narrative and disregard evidence that does not support that narrative (Kassin, Reddy, & Tulloch, 1990; Oswald & Grosjean, 2004; Nickerson, 1998).

Taken together, the problem with mathematical models is that they are not consistent with the way people generally think about, experience, and remember an event: as a narrative. Therefore, perhaps jurors are more likely to use explanation-based models rather than mathematical models to determine the story of what happened during a trial. However, that does not necessarily mean that people do not, or do not need to, use mathematical reasoning during a trial. For example, forensic evidence is often presented as likelihood ratios that jurors must interpret (National Institute of Forensic Science, 2017).

To understand how jurors' evaluate numerical differences in lineup evidence—the key manipulation in this thesis—we first must understand what lineup evidence is and how lineups are constructed. Much of the lineup literature has examined witnesses' memory and decision-making processes—which has led to reforms in the construction and administration of lineups. Although the primary interest of this thesis is juror decision-making with respect to lineup evidence, it is obvious to expect outcomes of witness decision-making would influence juror decision-making. Furthermore, based on the sheer magnitude of witness research in the lineup literature, it would be remiss to not provide at least an overview of the literature on witness decision-making to give historical context on the development and ongoing problems with lineup identification evidence.

Lineup Evidence

A lineup is a method of verifying that a suspect committed a crime. In its simplest form, a lineup consists of the presentation of a suspect along with similar-looking people who are known to be innocent (called *foils*). Witnesses are asked to view the lineup and identify if the perpetrator of the crime is present. A *physical lineup*—the type used in Beck's case—involves the suspect and foils presented in person standing in a line (hence

the term *lineup*) with witnesses viewing the lineup out of view of the suspect. A *photo lineup* involves witnesses viewing photographs of the suspect and a number of similar-looking foils. The lineup is used in most judicial systems around the world (Levi & Evans, 2008). Like the history of trial-by-juries, the exact origin of the lineup is unclear, however, the development of the lineup is strongly tied to the English judicial system: London police have been administering lineups at least since 1860 (Ackermann, 2017; Devlin 1976). The establishment and rise in use of lineups came out of the limitations of its much older predecessor: the *showup*. The showup was a simple procedure: After a crime was committed; police took the description of the perpetrator from the witnesses or victims; police investigated the crime and eventually identified a suspect; then police would show a photograph of the suspect to witnesses or victims and ask them if the person in the photograph was the perpetrator.

The showup is a simpler process than the lineup for police because it does not require finding appropriate foils, and therefore can be created and administered more quickly than lineups. The more rapid deployment of a showup can mean witnesses have better memories of the event as a result of less time for memory decay or interference, as well as the perpetrator having less time for changes to appearance (Neuschatz, Wetmore, Key, & Cash, 2016). Why, then, are showups no longer popular—and are in fact discouraged in many countries, including New Zealand (Evidence Act, 2006)? Researchers have argued that showups are problematic because they are suggestive of guilt. That is, witnesses often think the police have probably selected this person because they think the person is guilty, leading to increased chances of misidentifications (Levine & Tapp, 1973; Malpass & Devine, 1983). The showup also requires a dichotomous decision of either

"yes" or "no". If witnesses are uncertain and are merely guessing, then one would expect 50% of guesses to be "yes" responses implicating the pictured individual in the crime. It is also possible some of the "yes" responses are given by witnesses who are not very confident in the identification based on their memory, but they have faith that the police would identify the right person as the suspect, or because of their desire to punish someone for the crime (Malpass & Devine, 1983). The empirical evidence from research with witnesses viewing videos of simulated crimes demonstrates that showups lead to more false identifications than lineups (Gronlund et al., 2012; Lindsay, Pozzulo, Craig, Lee, & Corber, 1997). Indeed, many real cases of misidentification discovered through DNA evidence have been linked to showups (Garrett, 2011).

Although the literature suggests that lineups generally result in greater identification accuracy than showups, there are factors that can decrease the accuracy of lineups. Analyses of simulated crimes have shown that under certain conditions, lineups are no more accurate than showups (Gronlund et al., 2012). The way lineups are created and administered can influence the strategies witnesses use to make identification decisions. In the next section, I will describe the literature demonstrating how lineup construction and administration can affect witness behaviour. The purpose of this thesis was to take one of these influences on witness behaviour—an obvious lineup bias towards the suspect—and examine how jurors use this evidence to evaluate guilt. Although, it must be acknowledged that all of the outlined literature in the next section could be described as biasing witness behaviour, in this thesis we were specifically interested in biases towards the police suspect.

Lineup Construction

Unfortunately, a short summary on the literature on lineup construction is difficult and would be overly simplistic because many of the original findings and recommendations of the 1980s and 1990s have been challenged recently with improved methods and alternative statistical analyses (Wells et al., 2020). Furthermore, there are ongoing debates between researchers over the most appropriate recommendations for lineup construction and administration. This section provides an overview of the original findings and some of the recent challenges to these findings.

Foil selection. When police investigate a crime and find a suspect, they might conduct a lineup to verify the suspect's identity. But how should police select foils for their lineup? The answer to that question seems intuitive: you would select foils who look similar in appearance to the suspect. Indeed, according to New Zealand law, all members of the lineup should look similar and no one person should stand out (Evidence Act, 2006). However, researchers have long recognised that the foil selection is much more nuanced than selecting people who look similar in appearance (Lindsay & Wells, 1980; Luus & Wells, 1991). For one, precisely *how similar* should foils be to the suspect? Furthermore, on what features should police match the foils to the suspect? Instead of providing a definitive answer to these questions (which is beyond the scope of the current discussion), it is useful to instead think of the purpose of foils in a lineup. The primary purpose, of course, is to stop witnesses from using deductive reasoning to identify the suspect (Luus & Wells, 1991). That is, if witnesses cannot remember the perpetrator well, they cannot simply use the few details of the perpetrator they do remember to guess which member of the lineup most resembles that scant description of the perpetrator. Therefore, if foils are not similar enough to the suspect, witnesses who cannot remember the perpetrator are

likely to identify the suspect using deduction. However, if the foils look too similar to the suspect, then even witnesses with good memories of the perpetrator might mistakenly identify a foil. In other words, there is an ideal range in which foils are neither too similar nor too different from the suspect.

Luus and Wells (1991) proposed two foil selection strategies: a) matching the foils to the appearance of the police suspect; or b) matching the foils to features described in witnesses' descriptions of the perpetrator. The researchers reasoned that the *match-to-description* strategy would lead to all foils matching features that the witnesses had recalled in their police descriptions. The match-to-description strategy would also permit variation between people in the lineup in features not mentioned in the police description, which would result in enough variation between people in the lineup, making the recognition task easier for witnesses (i.e., increasing hits). Luus and Wells contrasted this approach to the *match-to-appearance* strategy, which they claimed could lead to people in the lineup being too similar in appearance, and therefore making the recognition task impossible (i.e., increasing misses and false alarms).

The first empirical test of these two strategies supported Luus and Wells (1991) predictions. False identification rates for both strategies were similar, but the rate of correct identifications was lower when foils were matched to the appearance of the suspect than when the foils were matched to witnesses' descriptions of the perpetrator (Wells, Rydell, & Seelau, 1993). However, subsequent research on these two methods has had mixed findings (see Fitzgerald, Oriet, & Price, 2015 for a summary). Meta-analyses comparing these two strategies have found no evidence that witnesses' decisions are more diagnostic

of guilt or innocence under either strategy (Clark & Godfrey, 2009; Clark, Howell, & Davey, 2008).

The lack of evidence to endorse either method does not mean, however, that foil selection is irrelevant. Fitzgerald, Price, Oriet, and Charman (2013) conducted a meta-analysis using studies comparing the similarity of foils in the lineup to suspects. The researchers coded lineups as being low, moderate, or high in suspect-foil similarity, based on independent subjective similarity ratings. The researchers found that suspect identification was more likely in lineups in which suspects were low in similarity to the foils compared with lineups in which the suspect was moderate or high in similarity to the foil. This finding was true when the culprit was present or absent. Furthermore, identification of foils was more likely in high and moderate similarity lineups compared to low similarity lineups. Differences between the low, moderate, and high similarity lineups did not lead to differences in people rejecting to make an identification. As the similarity between people in the lineups increased, the more likely a suspect identification was diagnostic of guilt. Therefore, high suspect-foil similarity leads to improved witness accuracy.

Simultaneous and sequential lineups. Once a lineup has been created, the way police administer the lineup to witnesses can also influence witnesses' behaviour. The idea that witnesses tend to pick the person from the lineup who looks most like the perpetrator instead of comparing each person to their memories was recognised in early psychological research on lineups (Wells, 1984). To eliminate these *relative judgments* without imposing a significant amount of extra work on police conducting lineups, Lindsay and Wells (1985) proposed and tested the *sequential lineup* method. A sequential lineup involves presenting

witnesses each lineup member separately, and the witnesses making a yes/no decision for each person. Importantly, to eliminate relative judgments, witnesses must make these yes/no decisions during their first viewing of the lineup. Also, to avoid witnesses comparing the last member of the lineup to the preceding lineup members, Lindsay and Wells also suggested that witnesses should not be informed of the number of people in the sequential lineup. Lindsay and Wells tested their sequential lineup by showing subjects in their experiment a staged crime. They found subjects in both simultaneous and sequential lineups were similarly accurate at identifying a guilty perpetrator. However, subjects viewing the simultaneous lineup were more likely than subjects viewing the sequential lineup to identify a person who resembled the perpetrator when the actual perpetrator was not in the lineup. That is, witnesses who viewed sequential lineups made fewer inaccurate decisions.

The superiority of sequential lineups over simultaneous lineups has been replicated many times (see Steblay, Dysart, Fulero, & Lindsay, 2001 and Steblay, Dysart, & Wells, 2011, for reviews). These experiments typically find that sequential lineups lead to a reduction in correct identifications and incorrect identifications—but the reduction in incorrect identifications is often greater than the reduction in correct identifications (Steblay et al., 2011). This shift in both correct and incorrect identifications for sequential lineups could mean people are generally more conservative when viewing a sequential lineup compared to a simultaneous lineup, rather than a superiority effect *per se* (Clark, 2012; Gronlund et al., 2014; Mickes et al., 2012). Indeed, recent research has questioned if sequential lineups are actually superior to simultaneous lineups. Early lineup research used *diagnosticity ratios*, the ratio of hits to false alarms, to show sequential superiority

(Lindsay & Wells, 1985). More recent research has advocated for the use of *discriminability,* based on *signal detection theory* (Egan, 1958; Green & Swets, 1966; Mickes, Flowe, & Wixted, 2012), which separates each persons' hit and false alarm rates to calculate a response criterion and discriminability. Some researchers claim the most appropriate method of displaying discriminability in the context of lineups is by using *Receiver Operating Characteristic* (ROC) curves, although the use of ROC curves over statistical methods is still subject to ongoing debate (Smith, Wells, Lindsay, & Penrod, 2017; Wixted, Vul, Mickes, & Wilson, 2018; Wixted, & Mickes, 2018).

Another important aspect of comparing sequential and simultaneous lineups is the consideration of order effects. What effect does the placement of the suspect and foil in the lineup sequence have on witnesses' judgments? There is research to suggest that witnesses' response criteria can change during a sequential lineup procedure; witnesses tend to be more conservative at the start of the procedure, and more liberal towards the end (Meisters, Diedenhofen, & Musch, 2018). Clark and Davey (2005) found that a foil who looked similar to the perpetrator was more likely to be incorrectly identified by witnesses when the foil appeared in the fourth position compared to the second position. In addition, when the perpetrator was present in a lineup, witnesses were more likely to correctly identify the perpetrator in the fourth position compared to the first position in the sequence (Meisters, Diedenhofen, & Musch, 2018). When position effects are controlled for, ROC curves for sequential and simultaneous lineups show similar rates of hits and false alarms. Therefore, there are strengths and weaknesses associated with using either simultaneous or sequential lineups—and the reported superiority of sequential lineups seems to have been premature because more recent research does not support this conclusion.

Biased Lineups

The final aspect of lineup construction that influences witnesses' behaviour—and most relevant for the experiments presented in this thesis—is *lineup bias*. It is useful to acknowledge that the word "bias" has different meanings in different areas of psychological literature (Einhorn, Hogarth, & Klempner, 1977; Kahneman & Tversky, 1979) for example, cognitive biases and social psychological biases often refer to systematic distortions that are somehow different to some objective measure of reality (Haselton & Nettle, 2015). In signal detection theory, a bias is the increased probability of one binary response, regardless of the stimulus (Green & Swets, 1966). For example, the probability of a person responding "yes" to any question. However, in the lineup literature, the term *biased lineup* is often defined as the extent to which the suspect is distinctive from the other members of the lineup (Brigham, Meissner, & Wasserman, 1999; Mansour, Beaudry, Kalmet, Bertrand, & Lindsay, 2017; McQuiston & Malpass, 2002). According to this definition, the bias is described as a characteristic of the lineup itself, rather than a person's behaviour—which is inconsistent with other usages of the word "bias" in psychology. However, this definition is based on the empirical finding that people tend to identify distinctive people from lineups (Buckhout, 1974; Shapiro & Penrod, 1986). In other words, a biased lineup is biased because it is likely to affect the responding of the witness in an unwanted manner due to either poor construction or administration of the procedure. This definition of bias was consistent with the focus of my thesis because it is presumably the bias that led to Beck's wrongful conviction.

There are at least two broad ways for a lineup to be biased towards a suspect. The first class of bias involves the construction of the lineup itself, specifically in terms of

differences in the appearance of lineup members. For example, if a perpetrator is described as having dark hair and a dark-haired suspect is put in a lineup with blonde-haired foils, the lineup would be biased towards the suspect because witnesses would likely select the suspect more often than the foils—regardless of their memories of the crime. This type of bias is not a simple biased or unbiased binary. For example, perhaps in the previous example, instead of all of the foils having blonde hair, it could be that only two foils have blonde hair and the remaining foils have black hair, in which case the lineup would be *less biased* than if all foils had blonde hair. That is, there are degrees of potential bias that can emerge based on the choices made during the construction of a lineup.

The second way for a lineup to be biased towards the suspect is by the other aspects of the lineup procedure, other than the similarity of the members of the lineup. That is, bias can be introduced in the administration of the lineup, including instructions and behaviours of the people conducting the lineup. An extreme (and hopefully extremely unlikely) example of this bias would be if the police told witnesses which member of the lineup was the suspect prior to asking them to actually identify the suspect. Another less obvious example of this type of bias is if witnesses view a lineup expecting the perpetrator to be present (Clark, 2005; Malpass & Devine, 1981). This expectation makes sense—it seems reasonable to believe that the police would not go through the hassle of assembling a lineup unless they, in fact, had a plausible suspect. This expectation, unfortunately, creates a bias for witnesses to identify *someone* from the lineup, even if they are not especially confident in their identification. To correct for this bias, it is common for police to tell witnesses that the "perpetrator may or may not be present." As part of this warning, sometimes witnesses are also told that it is just as important to free innocent people from

suspicion as it is to identify guilty people (Wogalter, Malpass, & McQuiston, 2004). This warning is referred to *unbiased lineup instructions.* Research on lineup instructions shows that witnesses are more likely to correctly reject to make an identification when the perpetrator is absent from a lineup under these unbiased instructions compared to no instructions.

Measuring lineup bias. As the previous sections demonstrate, there are many ways a lineup can be biased towards the suspect—and these different ways can vary in their degree of bias. The work on measuring lineup bias has exclusively focused on biases in the appearance of the lineup members (i.e., poor lineup construction)—to my knowledge there is no measure for the biases described in the previous section (i.e., poor lineup administration). There are two ways lineup construction fairness is typically measured. First, by estimating the number of plausible lineup members (Malpass, 1981; Malpass & Devine, 1983). The idea behind this measure is that if a lineup is biased such that some foils do not match the appearance of the perpetrator, they will not be counted as plausible alternatives to the suspect. In practice, the greater the number of easily excluded foils, the greater the bias toward the suspect.

The second, related, way of measuring lineup fairness is by measuring defendant bias—the extent to which witnesses select the suspect (defendant) over the other plausible member or members of a lineup. To calculate both types of lineup fairness measures, researchers use the *mock-witness procedure,* developed by Doob and Kirshenbaum (1973). The researchers had heard of a real case involving a perpetrator who was described by the witnesses as "rather good looking." It was reasoned that the witness might have simply identified the best-looking person from the lineup, regardless of whether that person was

the perpetrator or not. To test this possibility, the researchers gave subjects who had not witnessed the crime a description of the perpetrator and photograph of the lineup, and instructed them to identify the suspect. If these mock witnesses were able to correctly identify the suspect based on the description of the perpetrator alone, then the lineup was biased because the foils were not effective alternatives to the suspect. Actual witnesses might rely on their memories of the crime as well as other cues and strategies to make a lineup identification, (e.g., guessing or identifying the person who is the most distinctive). Mock witnesses can only use these other strategies because they do not have memories of the crime. Therefore, a lineup with a low bias in its construction should result in mock witnesses selecting a person from the lineup at chance. Doob and Kirshenbaum measured bias with this mock-witness procedure by calculating the proportion of mock witnesses who identified the suspect and used significance tests to compare that proportion to chance. Despite the development of this measure of defendant bias, Doob and Kirshenbaum's lasting contribution to the literature was their method of showing lineups to mock witnesses—a method which was used by other researchers to develop alternative ways of measuring lineup bias.

Functional size. One of these alternative measure of bias is *functional size*. This measure represents the total number of mock witnesses divided by the number of mock witnesses who identified the suspect (Wells, Leippe, & Ostrom, 1979). Functional size measures the number of people that are functionally present in the lineup. For example, in a six-person lineup, if there are ten mock witnesses and five identify the suspect, the functional size is two (10/5 = 2). Unlike Doob and Kirshenbaum's measure, functional size does not dependent on significance tests, and therefore the sample size of mock witnesses.

Functional size is, however, more reliable as the number of mock witnesses increases. However, functional size is also not a perfect measure, as Malpass (1981) points out. The functional size and the actual size of a lineup can be identical, yet some foils are less likely to be identified than other foils. Furthermore, in the case that no mock witness selects the suspect, functional size can actually be higher than the actual size of a lineup (Tredoux, 1988).

Effective size and Tredoux's E'. To overcome some of the problems with functional size, Malpass (1981) introduced effective size, which is similar to functional size, but takes into account the probability of mock-witnesses selecting the suspect by chance, and also takes into account the number of members of the lineup selected by mock witnesses. However, effective size does not take into account the number of null foils—foils who are not selected by any mock witnesses. Tredoux (1998) pointed out that null foils do occur and reformulated the effective size equation to consider that fact, as well as computing a formula that has a known sampling distribution. This improved measure is referred to E' or *Tredoux's E'*.

Jurors Evaluating Lineup Evidence

The evidence described in the previous section shows that biased lineups can clearly be problematic for witnesses' decisions: The greater the lineup bias, the more likely witnesses will identify the suspect, regardless of the suspect's actual guilt. One of the central foci of the current thesis involves examining how biased lineups influence the decision-making of individuals. It is important to acknowledge some ecological considerations before going further: In the real world, if a lineup is clearly biased, it is likely that the lineup evidence would be ruled inadmissible in a court of law. A biased

lineup would not constitute a *formal procedure* in New Zealand under the Evidence Act (2006), and without good reason to follow a formal procedure, prosecutors must prove the evidence is reliable. The judge would then rule whether the identification evidence is admissible or not. So, perhaps some biased lineups would be ruled inadmissible, and therefore might not influence jurors. However, the guidelines under the Evidence Act (2006) regarding lineup construction are not detailed enough to capture all lineups with some degree of bias. Specifically, Section 45 (3) (b) is the only specific reference to the actual construction of the lineup: "…in which the suspect is compared to no fewer than seven other persons who are similar in appearance to the suspect…" (p. 34). Without a specific measure of lineup fairness, as outlined in the previous section, it is possible that a lineup contains eight people who are similar in appearance, yet the lineup is nevertheless biased towards the suspect.

Therefore, it is at least plausible that evidence from a biased lineup identification could be presented before a jury in New Zealand or elsewhere. In fact, given the prevalence of faulty identifications in wrongful convictions, it seems quite likely that at least some of these identifications are a result of bias in the construction or administration of the lineup used. How, then, might jurors evaluate evidence from a biased lineup? According to the models of juror decision-making outlined previously, jurors should construct narratives of what happened using the evidence that is presented. If biased lineup evidence fits with jurors' narrative of what happened, then jurors should put more emphasis on that evidence. It is possible jurors might reject biased evidence, even if that evidence does fit with their narrative. In order for jurors to reject biased evidence, they must have the ability to identify a biased lineup. Devenport, Stinson, Cutler and Kravitz

(2002) showed jurors either biased or non-biased photo lineups and asked them to make multiple ratings. Jurors judged the biased lineup as less fair than the unbiased lineup and they judged the defendant as more culpable in the unbiased lineup compared to the biased lineup. Additionally, Wykes (2014) found jurors were more likely to convict a defendant when the witness made an identification from an unbiased lineup compared to a biased lineup. Therefore, it is likely jurors put less weight on biased evidence compared to unbiased evidence. However, there is no evidence in the literature to suggest jurors would completely reject biased evidence—as if it had not been presented at all.

Jurors Evaluating Biased Evidence from Multiple Witnesses

Taken together, the research suggests jurors are capable of identifying biased lineups, and that they are sceptical of biased lineup evidence. Recall, in the Adolf Beck case, although the lineups presented to witnesses were biased, the judge specifically commented about the overwhelming number of identifications. In other words, one person could have a poor memory or might make a mistake in identifying a suspect, but surely that many people could not be similarly mistaken. Should we expect, as in the Beck case, that jurors will rely on biased lineup evidence if there is an overwhelming number of unanimous witnesses identifying the suspect? To my knowledge, there is no published literature investigating this specific question. However, I will draw on research to predict how the number of witnesses might influence juror decision-making.

Before describing this literature, it is useful to determine how jurors *should* respond to biased lineup evidence. Should jurors be sceptical of biased lineup evidence regardless of the number of witnesses identifying the suspect, or does the number of witnesses identifying the suspect overcome some of the unreliability of biased evidence? This

question was addressed by Gunn et al. (2016) using mathematical models to estimate the probability of guilt of a person identified from a biased lineup by varying numbers of witnesses. In their models, they arbitrarily determined that there was a 50% chance that the suspect who appeared in the lineup was the true perpetrator. In addition, when the perpetrator was present, the researchers assumed witnesses accurately identified the perpetrator 52% of the time and failed to identify the perpetrator 48% of the time—these percentages were based off an estimate from previous empirical research (Foster, Libkuman, Schooler, & Loftus, 1994). To model the effects of a biased lineup, the researchers then added a small probability (0.01, 0.001, or 0.0001) of a systematic bias, such that the witnesses identified the suspect 90% of the time, regardless of whether that suspect was the perpetrator or not. Based on those assumptions, after between five to ten unanimous witnesses identifying the suspect (depending on the probability of the systematic bias), having additional unanimous witnesses identifying the suspect actually *decreases* the probability that the suspect is guilty.

Why does a greater number of unanimous witnesses lead to a lower probability of guilt compared to fewer witnesses? Memory is unreliable and a certain degree of inconsistency between witnesses should be expected. Furthermore, this inconsistency should increase as the number of witnesses increases. For example, if twenty people witness a crime, some of those witnesses should have a poor memory, which might cause them to identify a foil from the lineup. But if those twenty witnesses unanimously identified the suspect from the lineup, that could indicate that all twenty witnesses had a good memory of the event, they remembered the perpetrator well, and therefore all identified the suspect. Twenty unanimous witnesses could also indicate the witnesses were

influenced by factors other than their memory of the event, for example, some systematic bias in the way the lineup was conducted (Gunn et al., 2016). As the number of unanimous witnesses identifying the suspect increases, the less likely that unanimity is caused by a good memory, and more likely that unanimity was caused by a bias.

The pattern described by these mathematical models seem counterintuitive: rather than increased identifications overcoming the unreliability of a biased lineup, more unanimous identifications of the suspect could be a result of a biased lineup. This finding fits with Adolf Beck's case—the overwhelming number of identifications was likely a result of the biased lineup composition. If the lineup was fairer, perhaps not all of the 27 women would have identified Beck, which could have prevented Beck's wrongful convictions. However, these models do not tell us how actual jurors determine guilt when they are presented with lineup evidence from multiple witnesses.

There is reason to think that evidence from a greater number of witnesses is more compelling to jurors than evidence from fewer witnesses. In one study, subjects watched a lengthy trial in which employees of a railroad company claimed that the company caused them similar work-related injuries (Horowitz & Bordens, 2000). There were either one, two, four, six, or ten employees who were plaintiffs, each with similar symptoms, but with different jobs within the company and in different geographical locations. After hearing the evidence, subjects were asked to determine how liable the company was for the employees' injuries and to determine how much money the employees should receive as damage awards. As the number of plaintiffs increased, subjects rated the company as more liable for the injuries: Subjects in the four-plaintiff condition thought the plaintiffs should receive greater damage awards than subjects in the one- and two- plaintiff conditions.

However, subjects in the six- and ten- plaintiff conditions assigned lower damage awards than the four-plaintiff condition—suggesting a non-linear relationship between number of plaintiffs and damage awards. This finding was consistent with the authors' earlier study, using a different case, that found more guilty verdicts for subjects who were told there were hundreds of plaintiffs compared to subjects who were told there were 26 plaintiffs (Horowitz & Bordens, 1988). Therefore, the findings from these two studies support the idea that multiple people in agreement are more compelling than a single person, or two people, claiming the same thing. However, in these experiments, the identity of the defendant was undisputed—so the pattern of results observed might be different than the pattern of results observed for multiple witnesses identifying the same person from a lineup.

Taken together, the literature on multiple witness testimony shows that jurors should put more faith in witnesses' evidence as the number of consistent testimonies increases. However, to my knowledge, there is no published empirical research examining the effects of multiple witnesses' testimonies on jurors for biased lineups. Will jurors use the number of agreeing witnesses as an indicator of truthfulness, or will jurors completely disregard biased lineup evidence—regardless of the number of unanimous witnesses? Adolf Beck's case gives anecdotal evidence for the idea that at least judges might put more trust in biased lineup evidence if there is a large number of unanimous witnesses identifying the suspect. In Chapter 2, we tested this possibility. It is clearly a problematic finding if jurors do trust biased evidence, as demonstrated in Beck's wrongful convictions. If that is what we find, how might we reduce this reliance on biased evidence? In the next

section, I consider the potential effectiveness of a judicial warning on reducing juror reliance on identifications from a biased lineup.

Warnings

Recall, to overcome the expectation by witnesses that the perpetrator is present in the lineup, police often warn witnesses that the perpetrator may or may not be present, which increases the number of correct rejections. Could a warning have a similar effect to reduce reliance on biased lineup evidence? It is first important to consider the real-world applicability of the possibility of a warning in court. According to the New Zealand Evidence Act (2006), Section 122, if judges consider evidence *admissible*, yet *unreliable*, they can choose to caution jurors on accepting the evidence or advise them to consider the relative weight of that evidence. However, the specific wording of such a warning would be at the judges' discretion. Judges' ability to discriminate between reliable and unreliable witness evidence is, however, limited (Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006). Archival data from cases between 1980 – 2016 in Canada also show judges rarely discuss factors that influence eyewitness reliability in court (Bruer, Harvey, Adams, Price, 2017). Despite this research, it is at least possible that if biased lineup evidence is, for one reason or another, ruled as admissible evidence, and the judge is able to recognise the unreliability of the evidence, a judicial warning could be an option to reduce jurors' reliance on that evidence.

How might such a warning influence jurors? For a warning to have any influence on the decision-making process, jurors must comprehend the warning. Jurors do not always fully understand judicial warnings, but their comprehension of the implications of those warnings is improved when given in plain language, rather than in legal jargon (Gusick &

Ellsworth, 1992; Lieberman & Sales, 1997; Marder, 2015; Severance, Greene, & Loftus, 1984).

There are at least four possible effects that judicial warnings can have on juror decisions. First, judicial warnings may produce juror *confusion*, which is likely to result in either incorrect or at least inconsistent application of the warning to evidence (Martire & Kemp, 2009). Second, a judicial warning might result in increased juror *scepticism*, manifesting itself as a general disbelief in all of the evidence associated with the particular warning, regardless of the quality of the evidence (Cutler, Dexter, & Penrod, 1989). The third possibility is that the warning produces greater juror *sensitivity* to the quality of the evidence by applying the knowledge contained in the warning to the decision process. Ideally, jurors will become sensitive to witness accuracy. That is, jurors would be more likely to believe witnesses who are, in fact, more accurate. However, a more easily measurable type of sensitivity is sensitivity to the content of the judicial warning. That is, jurors' decisions would align with the warning (Martire & Kemp, 2009, 2011). For example, a warning about the unreliability of biased lineup might make jurors reject all biased lineup evidence—however it is still possible that the biased lineup is accurate despite the bias. The fourth possible effect is a *null effect*, which is when jurors do not apply the warning to the case (Henderson & Levett, 2016). A null effect can be observed experimentally using the same case details, and comparing jurors given a judicial warning and jurors not given a warning. If the decisions do not differ between the two groups, then that can be taken as evidence that the judicial warning had no influence on the outcome.

There appears to be no published literature on judicial warnings about biased lineups, however there are multiple studies examining the effects of judicial warnings

about the general reliability of eyewitness evidence. The majority of these studies use the same warning—the *Telfaire direction,* a judicial warning first given in *United States v Telfaire* in 1972. The Telfaire direction is a relatively vague instruction for jurors to consider whether the witness had the capacity to adequately observe the offender; if witness identification was based on their own recollection; and to consider the credibility of the witnesses. The Telfaire direction also emphases reasonable doubt "If after examining the testimony, you have a reasonable doubt as to the accuracy of the identification, you must find the defendant not guilty". Research using the Telfaire direction has had mixed findings. Some studies have found juror confusion (Greene, 1988; Hoffheimer, 1989); others have found juror scepticism (Greene, 1988; Katzev & Wishart, 1985). One study found juror sensitivity when the judicial warning was given before evidence was presented (Ramirez, Zemba, & Geiselman, 1996). Given that the Telfaire direction is vague, it is not surprising that some studies found evidence of juror confusion. Also, the emphasis on reasonable doubt might explain why studies have found juror scepticism. Taken together, the research using the Telfaire direction is mixed. It is likely that the methodological differences between these studies explain the different findings.

To my knowledge, and consistent with Leverick (2014), only one study investigating judicial instructions on eyewitness accuracy has found a sensitivity effect. Subjects in this study read transcripts of a court case. Subjects were either told that two witnesses discussed the crime together or that the witnesses did not discuss the crime (Paterson, Anderson & Kemp, 2013). Subjects were then either given a warning about the negative effects of co-witnesses discussion, or a general warning. After reading the court transcripts and judicial instructions, subjects gave verdicts. The researchers found no

evidence of scepticism—which would have been evident if there were fewer guilty verdicts overall for those subjects who heard the co-witness discussion warning (regardless of whether they were in the condition in which the witnesses discussed the event). However, the researchers found some evidence for sensitisation: Subjects believed evidence from witnesses who admitted discussing the crime less than evidence from witnesses who did not discuss the crime.

Taken together, the literature on judicial warnings is small, and many of these studies use the Telfaire directions. These studies most often show juror scepticism. However, that is not to say that judicial instructions cannot induce sensitivity. Studies examining the effects of warnings given by expert witnesses show that warnings can induce juror sensitivity (Leverick, 2014; Matire & Kemp, 2011). It is likely, then, that an effectively worded judicial warning could also sensitise jurors. However, there has been no research on warnings about biased lineups, let alone research examining the factors that might make warnings more likely to induce sensitivity. In this thesis, we created such a warning and tested the effects of the warning on juror guilt judgments.

Overview of Experiments

Across seven experiments we examined the extent to which the number of witnesses who identified the same member of the lineup influenced jurors' ratings of guilt. In Experiments 1, 2, 4, 5, and 7, we told some subjects that the number of witnesses changed during the presentation of evidence. Specifically, after we told subjects that a certain number of witnesses identified the suspect, and subjects rated the extent to which they thought that suspect was guilty, we then told them that one day later, another group of witnesses came forward and also identified the suspect. In Experiments 1 and 2 we also

examined the effectiveness of warning subjects that the lineup the witnesses viewed was biased, making them more likely to choose the suspect. In Experiment 3 and 6, we examined the effect of number of witnesses between biased and non-biased lineup administrations. Finally, in Experiment 5 and 7 we examined the effect of decreasing the number of witnesses identifying the suspect during the presentation of evidence.

Hypotheses and Predictions

Between-subjects effects. Subjects in conditions in which there are a greater number of witnesses identifying the suspect will judge that suspect as guiltier than subjects in conditions in which there are fewer witnesses identifying the suspect. Across the experiments presented in Chapter 2, we tested this prediction for both clearly biased lineups and unbiased lineups. For biased lineups, if the magnitude of identification evidence somehow makes the fact that the evidence is biased more acceptable to subjects, then we should observe higher guilt ratings for conditions with more witnesses identifying the suspect. If, instead, subjects reject biased evidence, regardless of the magnitude of identification evidence, then we would expect similar guilt ratings across conditions that differed by the number of witnesses identifying the suspect.

Within-subjects effects. To more directly test this idea that increasing the magnitude of identification evidence can somewhat negate the fact that the lineup evidence was biased, we also manipulated the number of witnesses identifying the suspect within-subjects. If subjects trust biased evidence more when there are more unanimous witnesses, then they should increase their guilt ratings when we tell subjects the number of witnesses identifying the suspect increased. There is a large body of literature describing the influence of prior information or decisions on future decisions, referred to as *anchoring*

and adjustment (Christensen-Szalanski & Willham, 1991; Fischhoff, 1975; Mussweiler, Strack, & Pfeiffer, 2000; Tversky & Kahneman, 1974). In general, in ambiguous situations, people rely on previously presented information or their previous decisions—even if this information or those decisions are irrelevant to the future decision. Experimentally, these anchoring effects are demonstrated by presenting subjects with an irrelevant number and then asking an unrelated question (Jacowitz & Kahneman, 1995). But anchors can also be self-generated by the subjects. For example, when people are asked when George Washington was elected as president of the United States of America, most people immediately think of the anchor date of 1776—the year the United States declared independence. People know this date is incorrect and that the true answer to the question must be after that date. People adjust from their anchor date of 1776, but their adjustment from this anchor is often insufficient (Epley & Gilovich, 2001; Jacowitz & Kahneman, 1995). Anchoring and adjustment effects have also been found in the court room. Kahneman, Schkade, and Sunstein (1998) proposed a model explaining how jurors decide upon punitive damages using an anchoring and adjustment strategy. There has also been research investigating anchoring effects in judges' sentencing decisions. When a prosecutor demanded a certain sentence length, judges used that demand as an anchor (Englich & Mussweiler, 2001). Therefore, we might expect subjects who are asked to make multiple guilt ratings to remember and then adjust their guilt rating based on changes in the evidence. More specifically, if the number of unanimous witnesses identifying the same person from a lineup increases during the presentation of evidence, then subjects should take their first guilt ratings and adjust them upwards. Conversely, if the number of

unanimous witnesses identifying a person from a lineup decreases during the presentation of evidence, then subjects should take their first guilt ratings and adjust them downwards.

Warnings. If subjects do trust biased evidence more if there are a greater number of witnesses identifying the suspect, then how might a specific warning about the biased lineup influence guilt ratings? We would expect that an effective warning would show a between-subjects effect. That is, subjects who are warned will generally give lower guilt ratings than subjects who are not warned. We might also observe within-subjects effects of warnings. That is, the addition of extra witnesses should not increase subjects guilt ratings in the warned condition to the same extent as the non-warned condition—this would also be consistent with subject sensitivity.

Table 1.

Summary of Experiments and Variables

Exp.	Witness Conditions	Bias	Warning	Dependent Variables	Crime scenario	Other IVs
1	Two conditions: "2 then 12" and "10 then 12."	One person with a distinctive feature (left scar; right scar; black eye).	Either warned or not warned that the lineup was biased.	Probability of guilt (using a sliding scale from 0 to 100).	Cellphone snatching in parking lot	
2	Six witnesses for all subjects, but presented as either 1, 2, or 3 separate groups coming forward.	One person with a distinctive feature (left scar; right scar; black eye)	Either warned or not warned that the lineup was biased	Probability of guilt (using a sliding scale from 0 to 100)	Same as Experiment 1	Guilt ratings either made after each new group of witnesses came forward or after all 6 witnesses had come forward.
3	Between-subjects conditions with either 5, 10, 15, or 20 witnesses coming forward as one group.	All lineup members matched the description (no distinctive features). Subjects in E3a were told which person was the prime suspect.	No warning	Probability of guilt (using a sliding scale from 0 to 100) and Verdict (Guilty and Not Guilty)	Bank robbery	
4	Two seperate crimes, one in which there were two witnesses and the other in which there were six witnesses.	Subject told that the same police officer investigated both crimes and identified the suspect to witnesses.	No warning	Probability of guilt (using a sliding scale from 0 to 100) and Verdict (Guilty and Not Guilty)	Case 1: an assault in a downtown bar. Case 2: an assault in a parking lot	
5	Same crime. One group initially told there were 2 witnesses, the other told there were 6 witnesses. Both groups told there was an administrative error and the number of witnesses they were told was incorrect (was either 2 then 6 or 6 then 2).	No bias	No warning	Probability of guilt (using a sliding scale from 0 to 100) and Verdict (Guilty and Not Guilty)	Same as Experiment 3	
6	Between-subjects conditions with either 1, 2, 3, or 4 witnesses coming forward as one group	Subjects in E6a were told which person was the prime suspect	No warning	Probability of guilt (using a sliding scale from 0 to 100) and Verdict (Guilty and Not Guilty)	Same as Experiment 3	
7	Same crime. One group initially told there were 5 witnesses, the other told there were 15 witnesses. Both groups told there was an administrative error and the number of witnesses they were told was incorrect (was either 5 then 10 or 15 then 10)	No bias	No warning	Probability of guilt (using a sliding scale from 0 to 100) and Verdict (Guilty and Not Guilty)	Same as Experiment 3	

Note. Exp. indicates the Experiment number and IV stands for independent variables,

Chapter 2

Across seven experiments, we examined the extent to which the number of unanimous witnesses identifying the police suspect from a lineup—who was the defendant in the court proceedings—influenced subjects. Table 1 summarises the main differences between each experiment. We aimed to address three main questions across the seven experiments. First, how does the number of witnesses identifying a suspect from either a biased or non-biased lineup influence guilt ratings? Second, how does warning subjects that the lineup was biased influence guilt ratings? Third, how does changing the number of witnesses identifying a suspect during the experimental session influence subjects' guilt ratings? To answer these questions, we presented subjects with limited information about a crime (we used four different crimes across our seven experiments), including a brief description of what happened and asked them to determine guilt as if they were making a decision like a juror prior to jury deliberation. All subjects then read that between two and twenty witnesses (depending on the experiment) saw the crime and each witness identified a single person from a police lineup. Subjects then rated that person's guilt.

To establish between-subjects patterns in the number of witnesses on subjects' guilt ratings, in Experiment 3 and Experiment 6, the experimental session ended after subjects gave a single guilt rating. In the remaining five experiments, we were interested in how guilt ratings changed if subjects were given updated information about the same crime, or in Experiment 4, if subjects were given information about a different crime. In four of our experiments, we gave updated information about the crime by telling subjects that the number of unanimous witnesses who identified the suspect had either increased (Experiments 1, 2, 5, and 7), or decreased (Experiment 5 and Experiment 7).

Experiment 1

Method.

Subjects. A total of 578 subjects completed the experiment using Amazon Mechanical Turk (www.mturk.com)[2]. Of these 578 subjects, 64.2% identified as women; 34.9% identified as men; and 0.9% identified as neither men or women. Subjects were between 18 and 75 years of age (M_{age} = 37.46, SD_{age} = 12.70, $Median_{age}$ = 34.00), and

[2] The seven experiments presented in this thesis are an exploratory investigation of the influence of the number of unanimous witnesses on juror decision-making. Across these experiments, we aimed for cell sizes of 50 subjects. According to G*Power 3.1 software, this cell size would produce sufficient power (1 - β error probability) of 0.81 to detect the difference between two independent groups using a *t*-test if Cohen's *d* was 0.57 (Faul, Erdfelder, Lang, & Buchner, 2007). Therefore, these *t*-tests would be sufficiently powered to detect (most) medium and large effect sizes, according to Cohen's classifications (medium Cohen's *d* = 0.5; large Cohen's *d* = 0.8) (Cohen, 1965). It would require cell sizes of 394 for the same level of power to detect a small effect size (of Cohen *d* = 0.2) between two independent groups (Bakker, Hartgerink, Wicherts, and van der Maas, 2016). Due to resource constraints, using cell sizes of 394 would have limited the number of conditions and experiments we could run in this thesis. Fewer experiments would have reduced the number of methodological adaptations we could make during the course of research and the more reliance on the results of each individual experiment. Instead, we decided to use moderate cell sizes and to test our main research questions over a greater number of experiments, with varied materials and scenarios.

received 0.25 USD upon completion. This research was approved by the School of Psychology Human Ethics Committee at Victoria University of Wellington.

Design. We used a 2 x 2 x 3 x 2 mixed design, with warning (warning or no warning), witness condition (2 then 12 or 10 then 12), and distinctive feature (left scar, right scar, or black eye) as between subjects independent variables and guilt rating (Time 1 and Time 2) as a repeated measure. We used subjects' guilt ratings (from 0 -100%) as our dependent variable.

Procedure. Qualtrics survey software was used (Qualtrics, Provo, UT) to present the experiments in subjects' web browsers.

Subjects were asked to evaluate evidence from a description of a crime that involved a man snatching a cell phone from another man in a parking lot. Subjects were either told that two people or ten people witnessed the crime. Police then created a description of the perpetrator based on the witnesses' descriptions. Subjects were presented with the police description, which described the perpetrator's height, age, hair colour, and one of three distinctive features—either a scar on the left cheek, a scar on the right cheek,

or a black eye. For example, subjects who saw the lineup with the black eye were told the perpetrator was a young male, 5'10" tall, curly hair, and had a black eye.

Figure 1. The left scar lineup from Experiment 1. In this lineup, Person 4 had a distinctive scar on the left side of his face. All three lineups and experimental instructions are presented in Appendix A.

Subjects were told that police had identified a man who lived close to the parking lot as a prime suspect. Subjects were then shown a photo lineup created by police, which consisted of a photograph of the suspect and five similar looking people. The lineup photographs and crime scenario were taken from standardised pools of faces, developed by

Colloff (2016) (see also, Colloff, Wade, & Strange, 2016).[3] All six photographs matched the generic description of the perpetrator given by the witnesses—except for height, which was indistinguishable in all of the photographs. Only one photograph (the suspect) had the distinctive feature described by witnesses (Figure 1 shows one of the three lineups used, the other two lineups are presented in Appendix A along with the exact instructions presented to subjects).

Subjects were told that each witness saw the lineup independently and the witnesses were asked if the perpetrator was present. All of the witnesses (either two or ten, depending on the condition) identified the person with the distinctive feature from the lineup (e.g., Person 4 in Figure 1); subjects were then told that person was, in fact, the police suspect.

---

[3] The photographs used to construct the lineups used in this thesis were from pools of faces created by Colloff (2016). For each pool of faces, a target culprit was determined. The remaining faces for each pool were selected using the following process: A group of 18 subjects were asked to watch a film of the culprit and answer a series of questions about the person's physical appearance. The modal descriptions of physical appearance were then entered into the Florida Department of Corrections Inmate Database (http://www.dc.state.fl.us/AppCommon/) to retrieve 40 photos of people who matched the description. The selected foils were all facing towards the camera with a neutral facial expression. The images were set to a neutral background, the clothing was changed to plain black, and the images were changed to grey scale. The resolution was altered so that the photos matched as closely as possible to the culprit.

After viewing the lineup, approximately half (51.2%) of the subjects were warned about the lineup being biased. In these conditions, subjects saw the lineup again, accompanied by a larger picture of the police suspect, a red arrow pointing to the distinctive feature, and the following text:

> As you can see, the lineup is not fair because only [Person x] fits the police description. [Person x] is the only person in the lineup with a [scar or black eye]. If I took a random group of people and gave them the police description and lineup, it is likely these people would pick [Person x], even though they themselves did not actually witness the crime.

All subjects then rated the probability that the police suspect was guilty, using a sliding scale from 0-100 (with 0 and 100 numerical anchors). After making this rating, subjects also provided a written explanation of their guilt rating. To measure subjects' memories for the primary manipulation, subjects were asked how many witnesses told the police they saw the crime and how many of these witnesses identified the suspect from the lineup.

Subjects then read that additional witnesses came forward one day after the first group of witnesses. Specifically, in the condition in which there were two witnesses, another ten witnesses came forward (the *2 then 12* condition). In the condition in which there were ten witnesses, another two witnesses came forward (the *10 then 12* condition). That is, after the additional witnesses came forward, there were now twelve witnesses in both conditions. Subjects read that a police officer visited the additional witnesses, one at a time, and showed each witness the lineup of the six faces. Each witness was asked if the perpetrator was present, and all of the new witnesses identified the police suspect.

To capture the effects of changing the number of witnesses, subjects provided a second guilt rating, and again gave a written reason for that guilt rating. To measure subjects' memory of the manipulation, subjects reported how many witnesses saw the crime, and how many witnesses identified the suspect from the lineup.

Subjects indicated how fair they thought the lineup was on a seven-point Likert scale, from one (*completely unfair*) to seven (*completely fair*). Finally, we asked subjects questions to identify those who failed to comply with our experimental instructions. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question in each experiment.

Results.

Compliance checks. Before addressing our research questions, we addressed compliance: 27.0% of subjects who finished the experiment told us they did not comply with our experimental instructions. However, many of these experimental instructions were purposefully strict, for example, subjects failed these compliance instructions if they indicated that they did not maximise their web browsers during the experiment. Therefore, to avoid excluding almost one-third of the data because of strict experimental instructions, we instead tested the extent to which failure to follow the instructions led to differing patterns of responses. Responses from those subjects who failed the compliance checks did not change the overall pattern of results, so we did not exclude them from our dataset (see Table B1). See also Appendix C for the subsequent analyses excluding these subjects.

Attention checks. We also identified subjects who failed to recall the number of witnesses to the crime. All subjects were asked how many witnesses, in total, saw the crime and how many witnesses identified the police suspect from the lineup. These

questions were asked twice (four questions in total): after the first group of witnesses came forward at Time 1 and after the additional witnesses came forward at Time 2. As Table B2 shows, 36.2% of subjects answered at least one of these four questions incorrectly. It is possible some of these subjects were not paying sufficient attention to the details of the crime. It is also possible that subjects were influenced by the number of witnesses mentioned in the details of the crime but they could not recall that number when subsequently questioned. Therefore, excluding these subjects might have excluded valid responses. Furthermore, responses from these subjects who failed to recall at least one of these questions did not change the overall pattern of results, so we retained them in the dataset. See Appendix C for the subsequent analyses excluding these subjects.

Lineup fairness. We also analysed subjects' ratings of how fair they judged the lineup. As expected, subjects who were told the lineup was biased gave lower fairness ratings ($M = 3.40$, $SD = 2.02$) than those subjects who were not told the lineup was biased ($M = 5.20$, $SD = 1.89$), $F(1, 566) = 121.15$, $p < .001$, Partial $\eta^2 = .18$. There were no other significant main effects or interactions (all $p > .10$). All interactions and main effects are presented in Appendix C.

Primary analysis. To examine our research questions, we ran a 2 x 2 x 3 x 2 ANOVA, with warning (warning or no warning), witness condition (2 then 12 or 10 then 12) and distinctive feature (left scar, right scar, or black eye) as between-subjects independent variables and guilt rating (Time 1 and Time 2) as a repeated measure. The mean and standard deviation for each condition are presented in Table D1. We first used the ANOVA model to test all interactions. The four-way interaction was not significant. All three-way interactions were not significant except the Guilt Rating x Distinctive

Feature x Witness Condition interaction, $F(2, 566) = 6.48$, $p = .002$, Partial $\eta^2 = .02$. The Guilt Rating x Warning two-way interaction was also significant, $F(1, 566) = 4.97$, $p = .03$, Partial $\eta^2 = .01$, and so was the Guilt Rating x Witness Condition two-way interaction, $F(1, 566) = 9.22$, $p = .003$, Partial $\eta^2 = .02$. The other two-way interactions were not significant. All main effects and interactions are presented in Appendix C.

Within-subjects effect of adding witnesses. Our primary analysis also showed a significant main effect of guilt rating, $F(1, 566) = 315.60$, $p < .001$, Partial $\eta^2 = .36$, which I will describe first, before addressing the interactions. This main effect shows that subjects' Time 2 guilt ratings were significantly higher than their Time 1 guilt ratings for subjects in all three distinctive feature conditions: left scar (2 then 12), $t(96) = 7.83$, $p < .001$, Cohen's $d = 0.69$; left scar (10 then 12), $t(94) = 5.44$, $p < .001$, Cohen's $d = 0.36$; right scar (2 then 12), $t(92) = 9.59$, $p < .001$, Cohen's $d = 0.69$; right scar (10 then 12), $t(97) = 5.10$, $p < .001$, Cohen's $d = 0.25$; black eye (2 then 12), $t(97) = 6.90$, $p < .001$, Cohen's $d = 0.48$; black eye (10 then 12), $t(96) = 8.72$, $p < .001$, Cohen's $d = 0.59$. That is, the addition of further witnesses identifying the suspect led to increases in subjects' guilt ratings.

Number of witnesses. I now return to the significant Guilt Rating x Distinctive Feature x Witness Condition interaction to understand the extent to which these increases in guilt ratings with the addition of further witnesses interacted with our other variables. We ran three follow-up Guilt Rating x Witness Condition interactions for each distinctive feature. In all three distinctive feature conditions, the overall pattern was the same. The Guilt Rating x Witness Condition interaction was significant, yet follow-up tests did not show any other significant differences: left scar $F(1, 190) = 7.33$, $p = .01$, Partial $\eta^2 = .04$;

right scar $F(1, 189) = 17.36$, $p < .001$, Partial $\eta^2 = .08$; black eye $F(1, 189) = 17.36$, $p < .001$, Partial $\eta^2 = .08$. Subjects who were told there were ten witnesses at Time 1 gave similar guilt ratings to those subjects who were told there were two witnesses at Time 1, the differences between guilt ratings in these conditions were not significant: $t(190) = .95$, $p = .35$, Cohen's $d = 0.14$; $t(189) = .53$, $p = .60$, Cohen's $d = 0.08$); $t(193) = .51$, $p = .68$, Cohen's $d = 0.01$. When subsequent witnesses were added at Time 2, those in the 2 then 12 condition gave similar second guilt ratings than those in the 10 then 12 condition which were also not significantly different across all three conditions $t(190) = 1.12$, $p = .26$, Cohen's $d = 0.16$; $t(189) = 1.88$, $p = .06$, Cohen's $d = 0.27$; $t(193) = .25$, $p = .80$, Cohen's $d = 0.04$. Refer to table D1, for the specific cell means for each condition. Therefore, despite the significant interaction, follow-up tests found no evidence of statistically significant differences in the increases in guilt ratings from Time 1 to Time 2 based on whether subjects saw different lineups with different distinctive features, or based on whether subjects were told ten witnesses or two witnesses came forward initially.

Warning. I now turn to our second research question. To what extent did warning subjects that the lineup was biased decrease their guilt ratings? As Figure 2 shows, subjects who were warned about the biased lineup gave lower guilt ratings than subjects who were not warned, demonstrated by a statistically significant main effect of warning, $F(1, 566) = 91.22$, $p < .001$, Partial $\eta^2 = .14$. The effect of warning can be described as a large effect size according to Cohen's (1969, 1988), somewhat arbitrary, classifications. To further put this effect size into perspective, the effect of warning observed in this experiment was larger than what is typically found in the well-documented literature examining the effects of defendant race on juror decision-making (Mitchell, Haw, Pfeifer, & Meissner, 2005

Mazzella & Feingold, 1994; Sweeney & Haney, 1992). We also found a significant Warning x Guilt Rating interaction. Subjects who were warned about the lineup being biased increased their guilt ratings from Time 1 to Time 2 to a greater extent, ($M_{\text{Time 1}}$ = 57.30, $SD_{\text{Time 1}}$ = 22.81; $M_{\text{Time 2}}$ = 70.78, $SD_{\text{Time 2}}$ = 24.11), $t(295) = 12.61$, $p < .001$, Cohen's $d = 0.57$, than subjects who were not warned ($M_{\text{Time 1}}$ = 75.34, $SD_{\text{Time 1}}$ = 21.44; $M_{\text{Time 2}}$ = 85.78, $SD_{\text{Time 2}}$ = 20.99), $t(281) = 12.46$, $p < .001$, Cohen's $d = 0.49$. Put another way, although the warning reduced subjects guilt ratings, when those warned subjects were told about the additional witnesses coming forward, they increased their guilt ratings to a greater extent than subjects who did not receive a warning.

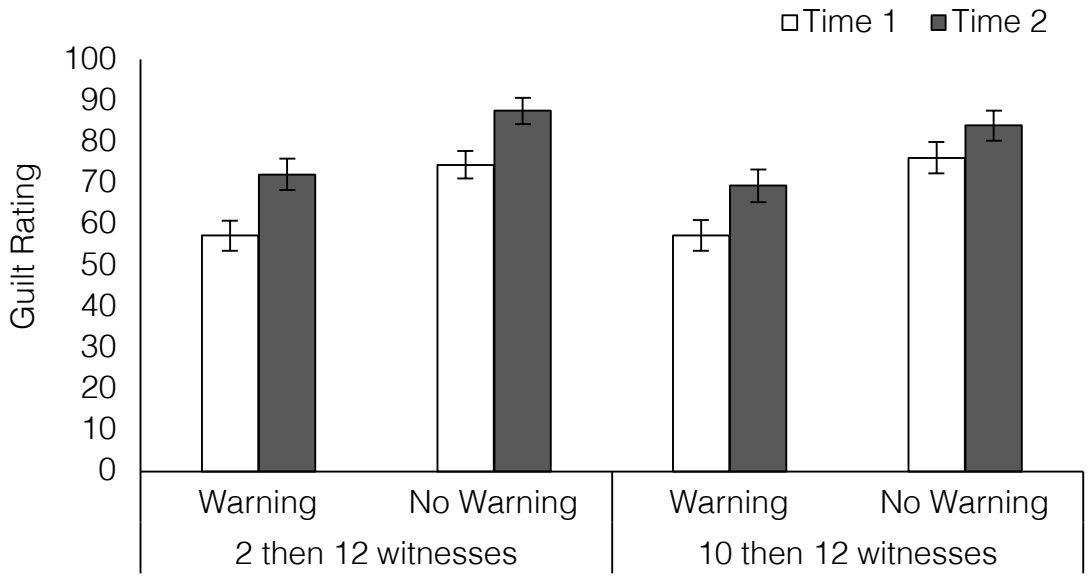

Figure 2. Mean guilt ratings for each condition in Experiment 1. The *Warning* bars represent those subjects who were warned that the lineup was biased and *No Warning* bars represent those subjects who were not warned that the lineup was biased. The light bars represent subjects' Time 1 guilt ratings and the dark bars represent subjects' Time 2 guilt ratings. Error bars represent 95% confidence intervals of the cell means.

It is possible the interaction with guilt rating and warning was caused by a ceiling effect. That is, subjects who were not warned gave higher guilt ratings at Time 1 and had less room to increase their guilt ratings further at Time 2. We tested this possibility by re-running the Warning x Guilt Rating interaction, but removing the 121 subjects who gave a guilt rating of 100 either at Time 1 or Time 2. With those subjects removed, we still observed a main effect of warning $F(1, 453) = 67.03$, $p > .001$, Partial $\eta^2 = .13$. However, unlike before, we found no evidence of a Guilt Rating x Warning interaction $F(1, 453) = 1.26$, $p = .26$, Partial $\eta^2 = .003$. That is, subjects who were warned increased their guilt ratings from Time 1 ($M = 53.02$, $SD = 20.41$) to Time 2 ($M = 65.82$, $SD = 22.60$) ($M_{diff} = 12.8$) to a similar extent to those subjects who were not warned: $M_{Time\ 1} = 69.28$, $SD_{Time\ 1} = 21.00$, $M_{Time\ 2} = 80.34$, $SD_{Time\ 2} = 22.42$ ($M_{diff} = 11.06$). Therefore, it is plausible that the Guilt Rating x Warning interaction observed in Figure 2 was caused by a ceiling effect. Indeed, that would be the most plausible explanation because we would not expect the effect of the warning to diminish over the short period between the presentation of the two groups of witnesses.

Guilt rating reasons. To get a richer understanding of subjects' decision-making during the experiment, we examined the explanations subjects gave for both guilt ratings. These explanations were individually examined and coded as unique reasons, which were determined by grouping together similar statements and statements with the same meaning. Most subjects gave multiple reasons for their guilt ratings, therefore the number of coded reasons was greater than the number of subjects. After the reasons were coded, the percentage of subjects who mentioned each reason was calculated. Table 2a shows the reasons subjects gave for their Time 1 guilt ratings, and

Table 2b shows the reasons subjects gave for their Time 2 guilt ratings. To highlight only key themes from subjects' explanations, these tables only display reasons that were mentioned by at least 5% of all subjects for each guilt rating. To highlight differences between the witness conditions—our primary interest—we also displayed these reasons by whether people were in the 2 then 12 condition or the 10 then 12 condition and whether they were warned about the lineup being biased.

As the Table 2a shows, a majority of subjects justified their Time 1 guilt rating because of the suspect's distinctive feature and resemblance to the description of the perpetrator (38.9% and 23.5%, respectively). The table also shows a small percentage of subjects (9.5%) specifically mentioned the number of witnesses who identified the suspect in their explanations. There was no significant difference between the percentage of people in

Table 2a

Reasons for Time 1 Guilt Ratings in Experiment 1

Reason	All subjects	Warning		No Warning	
		2 Witnesses	10 Witnesses	2 Witnesses	10 Witnesses
Mention of the distinctive feature	38.9	37.6	36.1	43.9	38.5
Uncertainty or there was not enough evidence presented	27.0	31.5	31.3	21.6	23.1
The person the witnesses identified matched the description of the perpetrator	23.5	17.4	25.2	26.6	25.2
Because witnesses identified the person	15.2	8.1	3.4	21.6	28.7

Because he was the police prime suspect	12.6	10.7	9.5	15.8	14.7
Mention the number of witnesses who identified the person	9.5	7.4	2.7	16.5	11.9
Because he lived close to where the crime was committed	7.4	12.1	9.5	4.3	3.5
because he has the distinctive feature	6.4	4.7	10.2	4.3	6.3
The probability he is guilty is 50/50	6.4	11.4	11.6	2.2	0.0
The distinctive feature makes the lineup biased	5.4	5.4	8.8	4.3	2.8
The lineup was biased	5.4	9.4	9.5	0.7	1.4

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 1 are listed.

the ten witnesses condition (7.2%) who mentioned the number of witnesses compared to the two witnesses condition, (11.8%) $Z = 1.89$, $p = .06$.

Table 2b shows the reasons subjects gave for their Time 2 guilt ratings. As the table shows, fewer reasons were mentioned by at least 5% of subjects for their Time 2 guilt ratings compared to subjects' Time 1 guilt ratings. The majority of subjects mentioned either the fact that witnesses identified the suspect, the fact that a greater number of witnesses identified the suspect since the last guilt rating, or they mentioned the specific number of witnesses who identified the suspect (31.0%, 33.9%, and 36.8%, respectively). Comparing, Table 2a and Table 2b shows more subjects mentioned the number of witnesses as a reason for their Time 2 guilt ratings compared to their Time 1 guilt ratings, $Z = 4.07$, $p < .001$.

Table 2b

Reasons for Time 2 Guilt ratings in Experiment 1

Reason	All subjects	Warning		No Warning	
		2 Witnesses	10 Witnesses	2 Witnesses	10 Witnesses
Mention the number of witnesses who identified the person	35.8	32.2	27.9	39.6	44.1
Because more witnesses identified the person	33.9	32.9	32.7	31.7	38.5
Because witnesses identified the person	31.0	23.5	28.6	43.2	29.4
Because the witnesses were unanimous	18.0	16.1	10.9	26.6	18.9
Uncertainty or there was not enough evidence presented	15.2	14.8	16.3	11.5	18.2
The distinctive feature makes the lineup biased	10.9	21.5	9.5	6.5	5.6
The same reason given for the first guilt rating	6.6	6.0	10.2	2.9	7.0

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 2 are listed.

Discussion.

We found mixed evidence for the effect of the number of witnesses identifying a suspect on subject's guilt ratings. On the one hand, we found no evidence for differences between guilt ratings at Time 1 when subjects were either told there were two or ten witnesses who identified the suspect. On the other hand, we did find evidence that the number of witnesses increased subjects guilt ratings when additional witnesses came

forward and identified the suspect. Below, I outline some of the possible explanations for these two findings.

There are at least two reasons why we found no evidence of differences in guilt ratings at Time 1. First, subjects might not have been specifically focused on the number of witnesses to determine their Time 1 guilt ratings. Subjects were likely establishing a narrative of what happened based on the information presented to them—consistent with the story model of juror decision-making (Pennington & Hastie, 1981, 1986, 1988, 1993). As part of this process, subjects would be integrating all of the evidence presented to create a coherent narrative. Indeed, the percentages on Table 2a show that subjects' their Time 1 guilt ratings were based on multiple different pieces of evidence. Perhaps, then, subjects were less focussed on the quality of each piece of evidence, and instead focussed on creating a narrative by integrating all of the evidence. Pennington and Hastie (1992) found that the ease at which subjects could create a narrative of what happened increased verdict judgments and perceptions of strength and credibility of evidence. In our experiment, at Time 1, being told there were two or ten witnesses probably did not make it easier for either group to construct a narrative of what happened, leading to similar guilt ratings. In other words, an increased number of witnesses did not add to the narrative value of the unanimous identification itself. However, once additional witnesses came forward, subjects' attention was drawn to the number of witnesses, which might have prompted them to think about how the number of witnesses influences their guilt ratings. If this explanation is true, then we should expect similar Time 1 guilt ratings between subjects, regardless of the number of witnesses subjects are told identified the suspect. For example,

if we added a condition with 25 witnesses, we would expect similar guilt ratings as observed in the two and ten conditions.

The second explanation for finding no evidence of differences in Time 1 guilt ratings could be because the manipulation of two versus ten witnesses was merely not sufficient to shift guilt ratings. If this was the case, then it would be inconsistent with Horowitz and Bordens (2000), who found subjects gave higher damage awards and liability judgments when there were ten plaintiffs compared to two plaintiffs. However, there are clear methodological differences between Experiment 1 and their experiment. Importantly, there was no lineup evidence presented in those experiments because the identity of the defendant was well established (the plaintiffs' employer). Additionally, an increase in the number of plaintiffs is plausibly more likely to have narrative value (e.g., ten injuries probably implies a greater level of negligence than two injuries). If, however, the manipulation used in our experiment was not strong enough to cause differences in guilt ratings, then we would expect that with larger differences in the number of witnesses coming forward (e.g., 25), we would eventually see differences in guilt ratings. We aimed to test these two explanations in the subsequent experiments.

Before addressing these two explanations, in Experiment 2, we aimed to address the within-subjects increases—the increase in guilt ratings between Time 1 and Time 2. Although it seems clear that this finding supports the idea that more witnesses increase subjects' certainty of guilt, it is also possible that guilt ratings are not dependent on the total number of witnesses coming forward, but rather the number of distinct groups of witnesses coming forward. This group-based explanation fits with the observed magnitude of increases between Time 1 and Time 2 in Experiment 1. Although in one set of

conditions, the number of witnesses increased by ten, there were no significant differences in these Time 2 guilt ratings compared to the set of conditions in which the number of witnesses increased by two. However, this comparison is not a direct test of this group-based explanation. A better test would be to vary the number of distinct groups of witnesses coming forward, while keeping the total number of witnesses equivalent across conditions. For example, we could present six witnesses coming forward as either three groups of two people, a group of two and a group of four, or one group of six people. We tested this possibility in Experiment 2.

There are at least two counter explanations that might have caused these increases in guilt ratings from Time 1 to Time 2, other than based on the number of witnesses or the number of groups of witnesses. First, the fact that some number of witnesses identified the suspect is repeated to subjects multiple times, and it could be that this repetition that increases guilt ratings, rather than the increase in the number of witnesses identifying the suspect. Indeed, Foster (2012) found that the number of times misinformation was repeated in eyewitness reports, not necessarily the number of different people attributed to those witness reports, increased subjects' belief in that misinformation. Similarly, we might expect that the number of times evidence is repeated increases guilt ratings, not the number of people associated with that evidence. One way to test this explanation is by simply repeating the witness evidence—and holding the number of witnesses constant. For example, telling subjects either once, twice, or three times, that two witnesses identified the suspect. However, we could not find a plausible reason for repeating evidence using our general methodology. Furthermore, the effects of repeating information have already been well-documented, and such an experiment would not address our main research questions.

It is, however, worth acknowledging that shifts in guilt ratings when additional witnesses come forward could be, at least in part, attributable to repetition effects.

The second counter explanation for these increases in guilt ratings is when additional witnesses come forward the act of giving multiple guilt ratings influences those ratings themselves. This idea fits with research that compared subjects who were asked to assess guilt either after each item of evidence was presented, or one global judgment of guilt after all evidence had been presented (Pennington Hastie, 1992). The researchers found subjects in the global judgment condition were more likely to use a narrative of what happened to determine guilt—consistent with the story model. In contrast, subjects in the item-by-item judgement condition were more likely to respond consistent with an anchoring and adjustment strategy to determine guilt—in which subjects create a numerical estimate of guilt based on the evidence, then adjust that numerical estimate as additional evidence is presented (Einhorn & Hogarth, 1985; Tversky & Kahneman, 1974). Perhaps subjects in our experiment used this anchoring and adjustment strategy when asked to give numerical estimates as additional evidence was presented. If subjects did not give these additional numerical estimates, then they might rely more on a narrative-based strategy to determine guilt. In that case, we would expect the number of additional witnesses identifying the suspect will have little influence on subjects' narratives of what happened. In Experiment 2, we tested this explanation by telling subjects that multiple groups of witnesses came forward at different times, and then asked some subjects to give one guilt rating after all witnesses had come forward and asked the other subjects to give guilt ratings after each new group of witnesses came forward. According to this anchoring adjustment explanation, we would expect those subjects who gave ratings after each group

of witness came forward would have a higher final rating than those subjects that only gave one rating.

Finally, before testing this counter explanation in Experiment 2, it is worth addressing the effect of warning subjects the lineup was biased. As expected, subjects who were warned the lineup was biased gave lower guilt ratings than subjects who were not warned. It is impossible to determine if this warning caused juror sensitivity or scepticism, which are discussed in the literature on warnings (Cutler, Dexter, & Penrod, 1989; Martire & Kemp, 2009; 2011; Katzev & Wishart, 1985; Ramirez, Zemba, & Geiselman, 1996; Greene, 1988). To test these two effects, we would need to give a general warning about biased lineups to subjects who saw a non-biased lineup and subjects who saw a biased lineup and compare guilt ratings. If subjects are sensitive to the warning, then the subjects who saw the non-biased lineup would give higher guilt ratings compared to subjects who saw the biased lineup. If subjects are sceptical, a warning should cause an equivalent reduction in guilt ratings for the biased and non-biased lineups. However, giving a warning in conditions with a non-biased lineup does not make practical sense in New Zealand. The warning used in this experiment informs subjects that the specific lineup used was biased towards the suspect. If New Zealand judges think evidence is unreliable, yet admissible, they can choose to warn subjects of the reliability of that evidence (Evidence Act, 2006). The warning we designed in this experiment was not intended as a general warning for all cases involving lineup evidence. Based on these circumstances, it is not problematic if the reduction in guilt ratings was due to sensitivity or scepticism. So long as the judge is able to recognise when to give the warning, then it does not matter whether jurors are sensitive

or sceptical. Put another way, it is more important that judges, rather than jurors, are sensitive to recognising biased and non-biased lineup evidence under these conditions.

Experiment 2

In Experiment 2 we examined the increases in guilt ratings between Time 1 and Time 2 that were observed in Experiment 1. Specifically, our primary research question was to what extent are these shifts due to the number of groups of witnesses coming forward at different times, rather than the total number of witnesses coming forward? Furthermore, does presenting the same number of witnesses coming forward and identifying the suspect as either one, two, or three groups influence subjects guilt ratings? Here, we use the word "group" to mean the number of independent witnesses coming forward at approximately the same time, not to indicate that these witnesses had any interaction with each other.

Our second research question was to test the anchoring and adjustment explanation that increases in guilt ratings are caused by subjects adjusting prior estimates of guilt. That is, if subjects give guilt ratings after each group of witnesses come forward, does that lead them to use an anchoring and adjustment strategy to determine guilt? In contrast, if subjects give one guilt rating after all witness evidence has been presented, does that lead them to use a narrative-based strategy for determining guilt? If the answer is *yes* to both of these questions, then we expect subjects who give guilt ratings for each new group of witnesses to have higher final guilt ratings than subjects who only give one guilt rating. We manipulated number of groups of witnesses while holding the actual number of witnesses who identified the suspect constant across all conditions.

Method.

Subjects. A total of 553 subjects completed the experiment using Amazon Mechanical Turk (www.mturk.com). We excluded five subjects who were not given information about the crime due to a technical error. Of the 548 remaining subjects, 59.5% identified as women; 40.5% identified as men. Subjects were between 18 and 85 years of age (M_{age} = 39.12, SD_{age} = 12.95, $Median_{age}$ = 36.00), and received 0.40 USD upon completion.

Design. We used a 3 x 2 x 2 x 3 incomplete factorial design, with number of witness groups (one, two, or three), warning (warned or not warned), guilt ratings (after each witness group or after all witness), and distinctive feature (left scar, right scar, or black eye) as between-subjects independent variables, and with final guilt rating as the dependent variable. Subjects in the one witness group conditions were asked to give guilt ratings after the first, and only, group of witnesses had come forward—which was also after all six witnesses had come forward. Therefore, the design was not completely factorial.

Procedure. We used Qualtrics survey software (Qualtrics, Provo, UT) to present instructions and questions in subjects' web browsers and TurkPrime to administer the experiment to subjects (Litman, Robinson, & Abberbock, 2017). Subjects completed the survey online via Mechanical Turk. To ensure subjects did not participate in multiple experiments, we retained records of subjects' Mechanical Turk worker identification numbers and set participation in a prior experiment (in this case, Experiment 1) as a disqualifying condition for each subsequent experiment. We used the same materials as Experiment 1, with two changes to the procedure.

The first change was to the total number of witnesses identifying the suspect. In all conditions, six witnesses identified the suspect from the lineup, but subjects read that these witnesses came forward as one, two, or three separate groups at separate times. Specifically, subjects in the three groups condition read that two people had witnessed the crime and identified the police suspect. Then one day later, two more witnesses came forward and identified the police suspect. Another day later, two more witnesses came forward and identified the police suspect. Subjects in the two groups condition first read that two people had witnessed the crime and identified the police suspect. Then these subjects read that one day later, four more witnesses came forward and identified the police suspect. In the last group, the one group condition, subjects read that six people had witnesses the crime and identified the police suspect. No other witnesses came forward in this condition. Thus, in all conditions, there were six total witnesses.

The second change was to how many times subjects gave guilt ratings. We asked approximately one third of subjects (37.2%)—only subjects in the two and three group conditions—to give guilt ratings after each time a new group of witnesses came forward and identified the suspect. Specifically, subjects in the three witness groups condition gave three guilt ratings and those subjects in the two witness groups condition gave two ratings. Clearly, this manipulation could not be used for those subjects in the one witness group condition because there were no additional groups of witnesses coming forward for them to make multiple judgments. Therefore, the design of the current experiment was not fully factorial.

Results.

Compliance checks. Before addressing our two main research questions we addressed compliance: 21.0% of subjects who finished the experiment did not comply with our experimental instructions. Responses from those subjects did not change the overall pattern of findings, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question. See also Appendix C for the subsequent analyses excluding these subjects.

Attention checks. We then determined which subjects failed to recall the number of witnesses who saw the crime. After each time subjects were asked to give a guilt rating, they were subsequently asked how many witnesses saw the crime and how many witnesses identified the police suspect from the lineup. As Table B2 shows, 30.1% of subjects answered at least one of these questions incorrectly. Responses from these subjects did not change the overall pattern of findings, so we retained them in the dataset. See Appendix C for the subsequent analyses excluding these subjects.

Lineup fairness. We also analysed subjects' ratings of lineup fairness. As expected, subjects who were warned the lineup was biased gave lower fairness ratings (M = 3.22, SD = 1.90) than those subjects who were not warned the lineup was biased (M = 5.08, SD = 1.86), $F(1, 530) = 131.91$, $p < .001$, Partial $\eta^2 = .20$. There were no other significant main effects or interactions (all $p > .1$). All main effects and interactions are presented in Appendix C.

Primary analysis. To answer our next questions, we ran a 3 x 3 x 2 ANOVA, with distinctive feature (left scar, right scar, or black eye), number of witness groups (one, two, or three), and warning (warned or not warned) as the independent variables. We did not include the guilt rating variable in this analysis because doing so would have violated the

assumption of independence between our independent variables: subjects in the one witness group condition were also those that gave guilt ratings both after each group of witnesses and after all witnesses had come forward because they only gave one guilt rating [5]. We used *final guilt rating* as the dependent variable for this ANOVA model, which was the rating subjects gave after being told all six witnesses identified the suspect, which was always the final guilt rating subjects gave in all conditions. We first ran all of the interaction terms and found none of them were significant. These interactions are listed in Appendix C. Also, see the means and standard deviations for each condition presented in Table D2. We used the main effects to answer our next four questions.

Distinctive feature. To check for differences between the three different lineups, we ran the distinctive feature main effect in our ANOVA model, $F(2, 530) = 0.39$, $p = .68$, Partial $\eta^2 = .001$, which was not significant. In Experiment 1, we observed a statistically significant three-way interaction involving the distinctive feature variable, but follow-up two-way interaction analyses showed no evidence of a different pattern of guilt ratings for the different distinctive features. Consistent with this finding, and as predicted, we found no evidence that using different faces and suspects with different distinctive features (shown in Appendix A) led to different final guilt ratings.

---

[5] To confirm the violation of the assumption of independence, we conducted a Pearson's correlation between the number of witness groups variable and guilt rating variable, which was large ($r = .63$) according to Cohen's (1988) guidelines. This correlation indicated there was not independence between our two variables (Mansfield & Helms, 1982).

Warning. Before addressing our two main research questions, we examined the extent to which warning subjects about the biased lineup reduced subjects' guilt ratings. To answer this question, we ran the warning main effect in our ANOVA model, $F(1, 530) = 57.54$, $p < .001$, Partial $\eta^2 = .10$, which was significant. Those who were not warned the lineup was biased gave higher final guilt ratings ($M = 83.41$, $SD = 19.28$) than those who were warned the lineup was biased ($M = 68.61$, $SD = 25.21$). This finding is consistent with Experiment 1, which found telling subjects the lineup was biased reduced their guilt ratings. The effect size observed was consistent with the large effect of warning observed in Experiment 1 (Partial $\eta^2 = .14$).

Number of witness groups. We then addressed our primary research question regarding the extent to which the number of groups of witnesses who identified the suspect from a lineup influenced guilt ratings. To examine this, we ran the number of witness groups main effect in our ANOVA model, $F(1, 530) = 5.84$, $p = .003$, Partial $\eta^2 > .02$, which was significant. Tukey post-hoc tests confirmed the pattern shown in Figure 3. Subjects in the three groups condition ($M = 76.38$, $SD = 24.79$) gave significantly higher final guilt ratings than subjects in the one group condition ($M = 69.66$, $SD = 24.79$) ($p = .028$, Cohen's $d = 0.27$). This effect of the number of groups is small according to Cohen (1988), and was comparatively smaller than the warning effect. Subjects in the two groups condition ($M = 77.98$, $SD = 21.71$) gave significantly higher final guilt ratings than subjects in the one group condition ($p = .004$, Cohen's $d = 0.36$). There was no evidence of significant differences between the three and two group conditions ($p = .74$). Therefore, we found some evidence that the number of groups of witnesses who identified the suspect influenced subjects' guilt ratings.

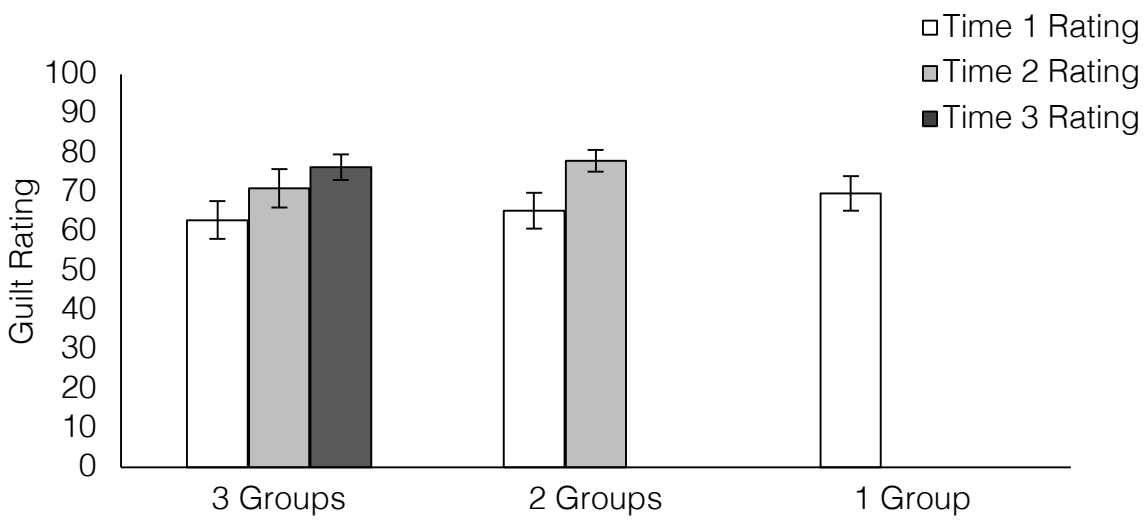

Figure 3. Mean guilt ratings for each witness condition in Experiment 2, collapsed across whether subjects were told the lineup was biased and whether they gave guilt ratings after each group identified the suspect or after all six witnesses identified the suspect. The light bars represent subjects' Time 1 guilt ratings, the grey bars represent subjects' Time 2 guilt ratings, and the dark bars represent subjects' Time 3 guilt ratings. Error bars represent 95% confidence intervals of the cell means.

Number of guilt ratings. We then addressed our second research question: To what extent does the number of times subjects gave guilt ratings influence those ratings? For this analysis, we excluded subjects in the one witness group condition. Excluding these subjects allowed us to re-run the primary analysis with the addition of the guilt rating variable, without violating the assumption of independence. Therefore, we ran a 3 x 2 x 2 x 2 ANOVA, with distinctive feature (left scar, right scar, or black eye), number of witness groups (two or three), warning (warned or not warned), and guilt rating (at the end or after each new group) as independent variables, with final guilt rating as the dependent variable. All four, three, and two-way interactions were non-significant (see Appendix C). We then

examined the guilt rating main effect which was non-significant. There was no evidence that guilt ratings differed based on whether subjects were asked to give a guilt rating after each new group of witnesses came forward (M = 76.60, SD = 25.30) or once after all of the six witnesses came forward (M = 77.73, SD = 21.31), $F(1, 411) = .001$, $p = .97$, Partial $\eta^2 < .001$.

Within-subjects shifts. Although, we were primarily interested in final guilt ratings, it was clear from Figure 3 that there was evidence of within-subjects shifts in guilt ratings when additional witnesses came forward—consistent with those shifts observed in Experiment 1. To statistically confirm this pattern, we ran three *t*-tests. First, we compared subjects in the condition with three groups who gave Time 1 and Time 2 guilt ratings: These subjects gave higher Time 2 ratings (M = 71.59, SD = 25.20) than their Time 1 ratings (M = 63.57, SD = 24.76), $t(137) = 6.00$, $p < .001$, Cohen's $d = 0.51$. Second, we compared subjects in the condition with three groups who gave Time 2 and Time 3 guilt ratings: These subjects gave similar—and not significantly different—guilt ratings, $t(137) = 1.59$, $p = .11$, Cohen's $d = 0.13$. Third, we compared subjects in the condition with two groups who gave Time 1 and Time 2 guilt ratings: These subjects gave higher Time 2 ratings (M = 77.70, SD = 22.77) than their Time 1 ratings (M = 64.08, SD = 24.59), $t(128) = 8.48$, $p < .001$, Cohen's $d = 0.75$. Therefore, consistent with Experiment 1, subjects shifted their guilt ratings upwards from Time 1 when additional witnesses came forward at Time 2.

We then examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes. As in Experiment 1, we included reasons that were mentioned by at least 5% of all subjects. The number of guilt ratings and

therefore reasons for guilt ratings differed by the condition in which subjects were assigned, we therefore concatenated each subject's reasons and coded them together—as one explanation per subject. We also collapsed across whether subjects made guilt ratings after each group of witnesses came forward or once after all the evidence had been presented. These reasons are presented on Table 3.

Subjects in the conditions in which the six witnesses came forward in two or three groups were more likely to justify their guilt rating with the witness evidence and they were more likely to mention the number of witnesses in their explanations than subjects in the conditions in which the six witnesses came forward all at once. Although the second-most common explanation was to reference the suspect's distinctive feature, it is likely that subjects were mentioning this to either justify low guilt ratings or high ratings. For example, subjects who simply wrote "because of the scar" might have either been referring to the fact that the suspect fits the description of the perpetrator and is therefore more likely to be guilty, or they might have been referring to the fact that the lineup was biased because the suspect was the only person with the distinctive feature and therefore less likely to be guilty. We know from the table that at least 14.1% of subjects mentioned the distinctive feature because it negatively affected the validity of the lineup. The responses also show a clear influence of the warning. Those subjects who were warned the lineup was biased were less likely to mention the witness evidence, express more uncertainty, and be more likely to mention that the suspect lived close to the location of the crime than subjects who were not warned. That is, subjects were less likely to mention the biased evidence and more likely to mention other evidence if they were warned, compared to the subjects who were not warned.

Table 3

Reasons for all Guilt ratings in Experiment 2

Reason	All Subjects	Warning			No Warning		
		3 groups	2 groups	1 group	3 groups	2 groups	1 group
Because witnesses identified the person	49.9	47.9	44.6	12.5	69.6	69.1	32.8
Mention of the distinctive feature	42.8	36.1	42.0	50.0	37.0	43.6	58.6
Mention the number of witnesses who identified the person	33.1	36.1	35.7	5.4	45.7	40.0	15.5
Uncertainty or there was not enough evidence presented	20.5	27.7	27.7	19.6	17.4	10.9	15.5
The person the witnesses identified matched the description of the perpetrator	19.0	16.8	18.8	23.2	12.0	20.9	27.6
Because the witnesses were unanimous	15.5	14.3	9.8	3.6	26.1	19.1	17.2
More witnesses came forward during experiment	15.0	20.2	15.2	0.0	20.7	20.0	0.0
The distinctive feature makes the lineup biased	14.1	18.5	24.1	10.7	13.0	7.3	3.4
Because he was the police prime suspect	8.8	5.9	6.3	16.1	4.3	12.7	12.1
The witnesses were independent	6.6	4.2	5.4	1.8	14.1	8.2	3.4
Because he lived close to where the crime was committed	6.2	5.9	10.7	17.9	2.2	0.9	3.4

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects are listed. Subjects gave one, two, or three explanations, depending on the conditions. For this analysis, we combined each subjects' explanations and coded them together. Reasons that were mentioned by the same subject in multiple explanations were only counted once.

Discussion.

The primary purpose of Experiment 2 was to further understand the within-subjects increases in subjects' guilt ratings when additional witnesses came forward and identified the suspect. When subjects were told that witnesses came forward in two or three smaller separate groups, that led to higher final guilt ratings than the condition in which six witnesses came forward as one group.

There are at least three possible explanations for this finding. First, we predicted based on Pennington and Hastie's (1992) research that subjects who were asked to give guilt ratings after each new group of witnesses would use an anchoring and adjustment strategy, and subjects who were asked for one guilt rating after all evidence had been presented would use a narrative-based strategy for determining guilt. Anchoring and adjustment would predict that the addition of witnesses would lead to increases in guilt ratings That is, each new instance leads to an adjustment upward from the previous judgment and thus, a higher and higher final guilt likelihood estimate. In contrast, the story model would not necessarily predict the addition of witnesses would lead to increases in guilt ratings. Despite these predictions, we found no evidence of significant differences between guilt ratings for subjects who gave guilt ratings after each group of witnesses came forward and those subjects who gave one guilt rating after all of the evidence had

been presented. Instead, we found evidence of this anchoring and adjustment strategy, regardless of the how many guilt ratings subjects gave.

The second reason why increases in the number of witness groups increased guilt ratings could be a result of how subjects conceptualised the witness evidence. We know that people often aggregate consistent and coherent information into one piece of evidence (Harris & Hahn, 2009; Pilditch, Hahn & Lagnado, 2018). That is, when witnesses hear multiple sources of evidence that are coherent, they create one representation of that evidence. In our experiment, the witness evidence is already presented in an aggregated form. However, when we presented witnesses as coming forward as distinct groups of witnesses, perhaps subjects did not aggregate that evidence and instead conceptualised the different groups of witnesses as two or three unique "pieces" of evidence—even though these pieces of evidence are consistent with each other.

A third reason why the number of groups of witnesses might influence guilt ratings in this experiment is the fact that *some* witnesses identified the suspect was repeated with each new group. That is, in the three witness groups condition, the suspect was identified by witnesses three times, compared to two times in the two witness groups condition and one time in the one witness group condition. Could this repetition, even in the absence of additional witnesses coming forward, influence guilt ratings? Foster et al. (2012) found the number of times information was repeated in evidence reports mattered more than the number of independent witnesses who repeated a claim. Therefore, it is likely that repetition, at least in part, explains the increases observed in guilt ratings when additional groups of witnesses come forward. In this thesis, we were not interested in further separating the effects of repetition from this manipulation because the effects of repetition

have been well documented (Mitchell & Zaragoza, 1996, Zaragoza & Mitchell, 1996; Unkelbach et al., 2007, Weaver et al., 2007). Furthermore, it is almost impossible to eliminate repetition as an influence if this was a real case in which groups of witnesses came forward at different times, and that evidence was presented as separate groups rather than aggregating the total number of witnesses. Therefore, it remains a possibility that any increases due to the number groups of witnesses coming forward could be at least in part, due to repetition of evidence.

One possible unintended consequence of the experimental design was the fact that the additional witnesses might have been perceived by subjects as being qualitatively different from those witnesses that came forth earlier. There is some evidence of this possibility in witnesses' reasons for their guilt ratings: Two subjects mentioned that these additional witnesses' memory of the crime and the perpetrator would not have been as accurate as those first witnesses. Although, not mentioned by any of the subjects, there are at least two other ways that the additional witnesses were qualitatively different from the first group of witnesses. First, the first group of witnesses might have put pressure on those additional witnesses to come forward to identify the suspect—even though we told subjects that the witnesses were independent. Second, the additional witnesses' description of the perpetrator was not used by police to create the lineup.

Although only two subjects mentioned that the additional witnesses might have been qualitatively different from the first group of witnesses, it is possible that other subjects thought so but did not write that reason in their written descriptions. One way to reduce the perceived memory differences is to have the additional groups of witnesses come forward hours after the first group, instead of one day later. However, if witnesses

came forward within hours, it would not make sense to present them as different groups. If we did present these witnesses as different groups, it would imply that the witnesses within each group came forward at the same time, or within minutes of each other—which would have made subjects suspicious and question their independence. Therefore, reducing the delay between groups of witnesses could result in a different concern for subjects. Furthermore, reducing the delay between groups does not preclude the idea that witnesses that came forward later were pressured by earlier witnesses. It is, therefore, unlikely that we can eliminate the fact that later witnesses coming forward might be perceived as qualitatively different to earlier witnesses. However, we should expect that subjects with these concerns are randomly assigned to conditions, so will have limited impact on the findings of our experiments (especially experiments in which the number of groups of witnesses is held constant, like Experiment 1).

Recall, we found no robust evidence of differences in guilt ratings based on the different lineups with different distinctive features. We further examined subjects' guilt ratings reasons and found 3.5% of subjects (all those in the black eye conditions) mentioned that if the perpetrator had a black eye at the time of the crime, then it would have healed by the time of the lineup one week later. That is, the black eye was a temporary distinctive feature—unlike the two scars. The fact that at least some subjects explicitly acknowledged this fact was reason enough to abandon the use of the black eye distinctive feature in future experiments.

Taken together, Experiment 2 has given further clarity on what might and might not explain the within-subjects shifts in guilt ratings when additional witnesses come forward. However, it still remains puzzling that we do not see differences between guilt ratings at

Time 1 based on the number of witnesses. This finding is inconsistent with Horowitz and Bordens (2002, 1988) and the much of the literature on how strength of evidence influences jurors (Devine, Olafson, Jarivs, Bott, Clayton & Wolfe 2004; Hannaford-Agor, Hans, Mott, & Munsterman, 2002). We wanted to be confident that the lack of difference between conditions in Time 1 guilt ratings was not simply a result of the specific methodology used in Experiments 1 and 2. There were two main methodological reasons why we might have not observed a difference in these guilt ratings. First, perhaps the number of witnesses does influence jurors' guilt ratings, but because we only used small differences of two or four witnesses between our conditions, we were unable to detect this effect. That is, the differences might simply be insufficient to produce a reliable effect. In Experiment 3, we addressed this issue, by increasing the differences in group size, by using group sizes of 5, 10, 15, and 20. Second, so far in Experiments 1 and 2, we have used biased lineups only. It might be that the initial lack of difference in guilt ratings is due to uncertainty caused by the biased lineup. Perhaps subjects disregard witness evidence because it is biased. In Experiment 3, we examined the extent to which using a biased lineup compared to a non-biased lineup influenced these Time 1 guilt ratings.

Experiment 3a and 3b

In Experiment 3a and 3b we addressed two questions. First, does the number of witnesses who identify the police suspect influence subjects' guilt ratings when all witnesses come forward at once? So far, in Experiments 1 and 2 we have presented subjects with groups ranging from two to ten witnesses. In Experiment 3, we extended this range (five to twenty) to examine the extent to which larger groups influence guilt ratings.

This extended range is more consistent with the range in which Gunn et al. (2016) calculated differences in diagnosticity for multiple unanimous witnesses.

The second question we addressed was to what extent do biased lineups influence subjects' guilt ratings compared to non-biased lineups? In Experiments 1 and 2, we only presented subjects with biased lineups. In Experiment 3 we either presented subjects with a lineup administration that was biased or one that was not biased. To avoid the effects of visual differences between conditions, the bias in the lineup used in Experiment 3 was not based on visual features of the lineup. That is, all faces in the lineup matched the description of the perpetrator (see Appendix A for these lineups). We manipulated bias in Experiment 3a by telling some subjects that a police officer told witnesses who the suspect was before the witnesses made a lineup decision (i.e., a bias in administration rather than construction). In Experiment 3b, we used the same lineup, but we did not tell subjects that a police officer identified the suspect. That is, there was no obvious bias in either lineup construction or administration for Experiment 3b[6].

In Experiment 3, in addition to making guilt ratings from 0 – 100, subjects also made a dichotomous *guilty* versus *not guilty* verdict choice. We added verdict because it is ultimately the decision jurors make in real cases. Additionally, we wanted to see if the differences observed in guilt ratings from our independent variables were similar to the pattern observed in verdicts. We predicted that both guilt ratings and verdict decision would yield the same pattern of results, but there was also the possibility that subjects

[6] We ran Experiment 3a before we ran Experiment 3b. Therefore, we used the naming convention 3a and 3b rather than running bias as an independent variable.

might be more conservative in verdict judgments compared to guilt ratings because verdicts have clear legal consequences (e.g., a fine, or imprisonment). In contrast, guilt ratings have no obvious legal consequences, so subjects might be more likely to give higher ratings. Of particular interest to us was the extent to which the number of witnesses identifying the suspect had a different effect on verdict judgments compared to guilt ratings.

We also made two minor changes to the appearance of the lineups in Experiment 3 compared to the previous experiments. First, to be consistent with New Zealand Police procedures, instead of using six-person lineups, we used eight-person lineups (Evidence Act, 2006). Second, instead of labelling the people in the lineup as numbers (e.g., Person 3), we labelled them as letters (e.g., Person C). We made this change to prevent the possibility that subjects were misremembering the name of the suspect (e.g., Person 3) as the number of witnesses who identified the suspect. These lineups are presented in Appendix A. Finally, we found no evidence in Experiment 1 and Experiment 2 that the different lineups with different distinctive features influenced guilt ratings, so in Experiment 3 subjects all saw the same lineup with no individuals within the lineup showing a distinctive feature.

Method.

Subjects. A total of 240 subjects in Experiment 3a and 242 in Experiment 3b completed the experiment using Amazon Mechanical Turk (www.mturk.com). Of these subjects, 38.3% in Experiment 3a and 43.0% in Experiment 3b identified as women; 61.7% in Experiment 3a and 56.6% in Experiment 3b identified as men. Subjects were between 19 and 71 years of age in Experiment 3a and 18 and 69 in Experiment 3b (Experiment 3a:

M_{age} = 35.54, SD_{age} = 10.56, $Median_{age}$ = 33.00. Experiment 3b: M_{age} = 33.94, SD_{age} = 9.54, $Median_{age}$ = 31.00). Subjects received 1.00 USD upon completion.

Design. We used a 4 x 2 design, with number of witnesses (5, 10, 15, or 20) and lineup administration (biased or not biased) as independent variables, and with guilt ratings as the dependent variable. The lineup administration variable was manipulated between Experiment 3a and 3b.

Procedure. Subjects were told that they would evaluate evidence from a bank robbery, involving a man walking into a busy bank and calmly approaching a bank teller and demanding money. The bank teller complied with the robber's instructions and handed him the money. Subjects were told there were either five, ten, fifteen, or twenty witnesses who saw the crime and gave a description of the perpetrator to police.

Subjects read the police description of the perpetrator, which included his height, age, hair, and ethnicity—unlike Experiment 1 and 2, the description matched all of the members in the lineup (except for height, which was indistinguishable between lineup members). After reading the description of the perpetrator, subjects read that police had "identified a man who lived close to the bank as a suspect, and considered him their prime suspect." Police then created a lineup using a photo of the suspect and seven similar looking people. Subjects then viewed the lineup that the witnesses had seen. Subjects read that each witness viewed the lineup, independently from each other, and was asked if the person who robbed the bank was in the lineup. In Experiment 3a, subjects read that prior to each witness making a lineup decision, a police officer told them that Person D was their prime suspect. This information was not given to subjects in Experiment 3b. Experiment 3a and 3b were otherwise identical. According to the information presented, all of the

witnesses (five, ten, fifteen, or twenty, depending on the condition) identified Person D as the perpetrator.

Subjects then gave guilt ratings, using a sliding scale (from 0-100), and gave written justifications for why they picked that rating. We then asked subjects if they were on a jury, what their verdict for the police suspect would be—guilty or not guilty. Unlike a real court proceeding, there was no subsequent jury deliberation.

Subjects then answered the same questions we asked in Experiment 1 and 2, to determine if they could recall the number of witnesses, and to rate, on seven-point scales, the extent to which they thought the lineup, and the administration of the lineup, were fair. Finally, subjects answered questions to identify those who failed to comply with our experimental instructions.

Results.

***Compliance checks*.** We first addressed compliance: 14.7% (12.1% in Experiment 3a and 17.4% in Experiment 3b) did not comply with experimental instructions. These responses did not change the overall patterns in subsequent analyses, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question. See Appendix C for the subsequent analyses excluding those subjects who failed the compliance questions.

Attention checks. We also determined which subjects did not recall the number of witnesses who identified the suspect: 16.6% (19.2% in Experiment 3a and 14.2% in Experiment 3b) failed at least one of the two questions (see Table B2). Responses from these subjects who failed to recall at least one of these questions did not change the overall

patterns in subsequent analyses, so we retained them in the dataset. The subsequent analyses excluding these subjects are also presented in Appendix C.

***Lineup fairness*.** We then examined subjects' ratings of lineup fairness and ratings of the fairness of the lineup administration. Subjects who were told the lineup administration was biased gave lower lineup fairness ratings (M = 4.42, SD = 2.23) and lower administration fairness ratings (M = 3.65, SD = 2.43) than those subjects who were not told the lineup administration was biased (lineup fairness: M = 6.05, SD = 1.21) (lineup administration fairness: M = 5.88, SD = 1.30), (lineup fairness: $F(1, 474) = 99.29$, $p < .001$, Partial $\eta^2 = .17$) (lineup administration fairness: $F(1, 474) = 157.25$, $p < .001$, Partial $\eta^2 = .25$). There were no other significant main effects or interactions (all $p > .48$). All of these non-significant main effects and interactions are presented in Appendix C.

***Number of witnesses*.** We addressed our primary research question: To what extent did the number of witnesses who identified the suspect influence subjects' guilt ratings? Although Experiment 3a and 3b were experiments conducted at different times, for the purpose of answering this question, we treated bias as an independent variable, so we could directly compare the influence of bias and non-biased lineups. We ran a 4 x 2 ANOVA, with number of witnesses (5, 10, 15, or 20) and lineup administration (biased or not biased) as independent variables, with guilt rating as the dependent variable. The means and standard deviations for each condition are presented in Table D3. The Number of Witnesses x Lineup Administration interaction was not significant, $F(3, 474) = 0.11$, $p = .96$, Partial $\eta^2 = .001$. As Figure 4 shows, the main effect of witnesses was also not significant, meaning we found no evidence that the number of witnesses influenced subjects' guilt ratings, $F(3, 474) = 0.39$, $p = .76$, Partial $\eta^2 = .002$. This finding is

consistent our findings in Experiment 1, showing no evidence of a between-subjects main effect of guilt ratings. Even with a greater range of witness group sizes, we found no evidence that the number of witnesses who identified a suspect from a lineup influenced jurors' guilt ratings. Furthermore, the non-significant Number of Witnesses x Lineup Administration interaction shows no evidence that biased and non-biased lineups influenced subjects guilt ratings differently when the number of witnesses identifying the suspect varied.

Bias. We then addressed our second research question: To what extent did subjects give different guilt ratings for biased and non-biased lineups? To answer this question, we used the 4 x 2 ANOVA reported above. Recall, the Number of Witnesses x Lineup Administration interaction was non-significant, so we analysed the main effect of lineup administration, which was significant, $F(1, 474) = 76.46$, $p < .001$, Partial $\eta^2 = .14$. As Figure 4 shows, those in Experiment 3a—the biased administration—gave lower guilt ratings ($M = 60.74$, $SD = 26.44$) than subjects in Experiment 3b—the non-biased administration ($M = 79.47$, $SD = 19.98$). Even in the absence of a warning, subjects used the information about bias to help determine their guilt ratings. The effect size (Partial $\eta^2 = .14$) of bias on guilt ratings was identical to the effect size of an explicit warning in Experiment 1.

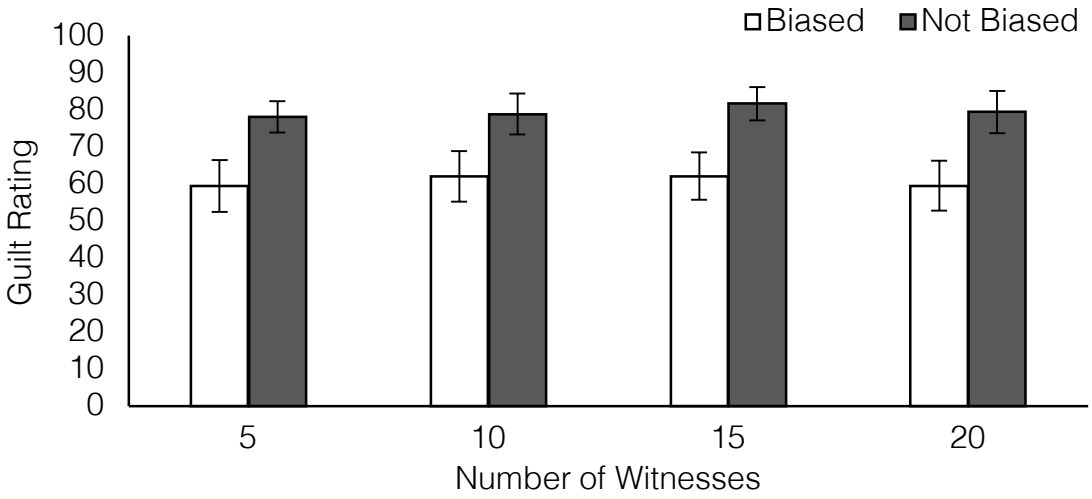

Figure 4. Mean guilt ratings for each witness condition in Experiment 3. The light bars represent subjects in Experiment 3a, who were told a police officer identified the suspect to witnesses (biased condition). The dark bars represent subjects in Experiment 3b, who were not given any information about how the police officer administered the lineup (not biased condition). Error bars represent 95% confidence intervals of the cell means.

Verdict. Finally, we re-ran our primary analysis using verdict as the dependent variable instead of guilt rating. We did this because in the explanations subjects gave for their guilt ratings in Experiments 1 and 2, some subjects had mentioned that although they had given a high guilt rating, they still had reasonable doubt and would not have given a guilty verdict. We tested this idea by using subjects' verdict responses as the dependent variable to re-answer our first two research questions. The mean guilt rating for those subjects who gave a guilty verdict was 81.95 (*SD* =16.12), compared to 45.94 (*SD* = 23.12) for subjects who gave a not guilty verdict. We ran a logistic regression with number of witnesses as the predictor variable and verdict as the outcome variable. The model was significant X^2 (4) = 55.84, $p < .001$. There was no evidence that the number of witnesses

influenced guilty verdicts (Wald = 1.79, $p = .62$). Consistent with the ANOVA showing a main effect of administration on subjects' guilt ratings, the model showed that fewer subjects in the biased experiment gave guilty verdicts (51.7%) than the non-biased experiment (82.6%) (Wald = 49.32, $p < .001$). That is, we found no evidence of differing patterns of results by using either guilt ratings or verdict as the dependent variable.

Finally, we examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes, and displaying the reasons mentioned by at least 5% of subjects on Table 4. There are three clear patterns on the table. First, fewer subjects in the conditions in which the lineup administration was biased mentioned the witness evidence or the specific number of witnesses in their explanations compared to subjects in the conditions in which the lineup administration was not biased ($Z = 8.12$, $p < .001$ and $Z = 4.76$, $p < .001$, respectively). This finding is consistent with Experiment 1 and Experiment 2 which both found that warning subjects of a biased lineup led to fewer subjects mentioning the witness evidence. In this experiment, none of the subjects received explicit warnings about the bias—those in the biased conditions were simply told that the person conducting the lineup identified the suspect to witnesses prior to the witnesses making a lineup decision. Many subjects in the biased conditions probably realised the problem with the evidence without an explicit warning. The second finding on Table 4 is that fewer subjects in the non-biased administration conditions mentioned the unanimity of witnesses in their explanations than subjects in the biased administration conditions ($Z = 7.20$, $p < .001$). Presumably, this difference was caused by some subjects in the biased condition realising that unanimity might be caused by the bias. However, even in the biased conditions, there was a large percentage of subjects who mentioned

unanimity. While coding these responses, I found many subjects mentioned that unanimity offset some of the problems associated with the biased lineup administration. For example, one subject wrote, "I think when the officers told the witnesses that Person D was their prime suspect it biased their thinking. They should have left that information out to get a more accurate judgement. Still, the fact that all 15 said it was Person D makes it fairly likely it was him." The third clear finding on Table 4 is that none of the subjects who were in the non-biased administration conditions mentioned that the lineup was biased or that police tainted the investigation. This pattern shows that our manipulation of bias worked and confirmed that the lineup materials we used were not in some way unintentionally biased. After all, we did not quantitatively measure bias in our experiments, so there was a possibility that the lineups we constructed were unintentionally biased. However, if there was a perceivable bias in the materials, we might expect some subjects in the non-biased conditions to mention a bias in their explanations.

Table 4

Reasons for Guilt ratings in Experiment 3

		Biased				Not Biased			
Reasons	All subjects	5	10	15	20	5	10	15	20
Because witnesses identified the suspect	61.2	46.7	36.2	45.9	42.6	75.4	82.0	79.3	79.0
Mention the number of witnesses who identified the person	33.8	20.0	20.7	27.9	24.6	42.6	44.3	43.1	45.2
Because the witnesses were unanimous	48.1	33.3	32.8	26.2	34.4	67.2	67.2	62.1	61.3

Police tainted/biased the evidence when they identified the suspect to witnesses	20.7	41.7	46.6	36.1	42.6	0.0	0.0	0.0	0.0
Uncertainty or there was not enough evidence presented	16.6	11.7	19.0	21.3	24.6	14.8	13.1	19.0	9.7
Eyewitness evidence or eyewitness memory is unreliable	8.1	5.0	8.6	6.6	4.9	11.5	4.9	10.3	12.9
Multiple people in the lineup matched the description	5.2	1.7	10.3	3.3	3.3	3.3	3.3	10.3	6.5

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects are listed

Discussion.

In Experiment 3 we manipulated lineup bias. To eliminate perceptual differences between conditions if we had used a lineup construction bias, we manipulated whether or not the police identified the suspect to subjects prior to subjects making a lineup decision (i.e., a lineup administration bias). The manipulation was successful: Subjects in the biased experiment rated the lineup administration as significantly less fair than the subjects in the non-biased experiment. Subjects in the biased experiment also rated the lineup itself less fair than subjects in the non-biased experiment—it is likely this finding was caused by subjects misunderstanding the question because the lineups were identical. The question about whether the lineup itself was biased was presented first before the question about the administration, and subjects likely thought the first question was asking about the lineup administration bias.

One of our primary questions in Experiment 3 was to what extent do the number of witnesses identifying the suspect influence guilt ratings, for both either biased and non-biased lineups. Recall, Gunn et al. (2016) found that the functions of diagnosticity of

multiple witnesses differed depending on the degree of bias in a lineup. We did not expect subjects to be as sensitive to changes in bias as these mathematical models, but it was possible that we might have observed between-subjects differences in guilt ratings when the number of witnesses varied for nonbiased lineups. That was not the pattern we observed. Although subjects in the biased experiment gave lower guilt ratings overall than subjects in the non-biased experiment, there was no evidence that the number of witnesses influenced those ratings in either experiment. These findings help us understand the Time 1 ratings in Experiment 1. Recall, we found no evidence of differences in guilt ratings between subjects who were told there were two witnesses compared to those who were told there were ten witnesses. The findings from Experiment 3 suggest the findings in Experiment 1 were unlikely to be caused by using only biased lineups nor due to the specific characteristics of the crime. It is also unlikely that the manipulation of witness numbers was not strong enough in Experiment 1 (a difference of eight witnesses between the two conditions). In Experiment 3 we found that even with a difference of 15 witnesses between conditions, we observed no evidence of differences between guilt ratings.

The findings of the current experiment support the explanation that subjects are engaged in a narrative creation process to determine what happened when they are presented with the witness evidence (Pennington & Hastie, 1981; 1986; 1988; 1993). It seems that the number of witnesses who identify the suspect does not make it easier for subjects to determine what happened. We would, however, expect if there were witnesses who each contributed different information, that would influence guilt ratings—because that different information might help subjects create a narrative of what happened. Indeed, there is published research examining whether multiple witnesses simply corroborate each

other, or whether they present different, but consistent, information—which shows that arguments presented as coming from multiple sources were more persuasive than the same arguments presented as coming from a single source. (Pickel, 1993; Harkins & Petty, 1981). However, such a manipulation was not within the scope of this research programme.

Now that Experiment 3 has given further confirmation and explanation for the lack of evidence of differences in initial guilt ratings, we turn to focus on explaining the within-subjects shifts in guilt ratings observed in Experiments 1 and 2. Recall, I suggested these shifts in guilt ratings might be explained by either an anchoring and adjustment strategy or by the repeated identification of the suspect. Of course, these two explanations are not mutually exclusive, the shifts could be explained by a combination of both. One way to further clarify the mechanisms behind these shifts is to eliminate the influence of one of these potential explanations. In Experiment 4, we aimed to eliminate repetition as a plausible explanation by asking subjects to simultaneously evaluate evidence from two different crimes. Instead of having additional witnesses come forward within the same cases—and therefore repeatedly identifying the same suspect—we presented similar cases, that varied in the number of witnesses.

Experiment 4

In Experiment 4 we aimed to test our anchoring and adjustment explanation for shifts in guilt ratings in the absence of repeatedly identifying the police suspect. We presented each subject with two crime scenarios and asked them to make guilt ratings about both crimes. In the first crime scenario, we would expect subjects to give similar first guilt ratings, regardless of differences in the initial number of witnesses—which would be

consistent with what we observed in Experiments 1 and 3. Then, if subjects are presented with a second crime and told a greater number of witnesses identified the suspect than in the first crime, they should give higher second guilt ratings than their first guilt ratings. This finding would be consistent with the anchoring and adjustment explanation because the first guilt rating should act as an anchor for the second guilt ratings, even though the ratings are about two different crimes. Indeed, anchoring an adjustment effects have been found when the anchoring number is unrelated to subsequent numerical estimates, such as a number spun on a roulette wheel influencing subjects answers to general knowledge questions. (Tversky & Kahneman, 1974). Additionally, if subjects are presented with a second crime and told fewer witnesses identified the suspect than in the first crime, according to the adjustment explanation they should give lower second guilt ratings than their first guilt ratings. This manipulation did not involve repeatedly identifying the suspect to subjects, so if the manipulation does not influence guilt ratings, then perhaps the within-subjects shifts in guilt ratings observed in our previous experiments can solely be attributed to repeatedly identifying the same suspect.

Method.

Subjects. A total of 204 subjects completed the experiment using Amazon Mechanical Turk (www.mturk.com). Of these 204 subjects, 51.0% identified as women and 49.0% identified as men. Subjects were between 18 and 75 years of age (M_{age} = 35.57, SD_{age} = 11.19, $Median_{age}$ = 33.00), and received 0.80 USD upon completion.

Design. We used a 2 x 2 mixed design, with order (2 then 6 or 6 then 2) and guilt rating (Case 1 rating and Case 2 rating) as independent variables, with order as a between-subjects independent variable and guilt rating as a repeated measure.

Procedure. Subjects evaluated evidence from two assault cases: Case 1 involved an assault in a downtown bar, and Case 2 involved an assault in a parking lot. Subjects read that the same police officer had investigated both assaults. In both cases, a man attacked another man and immediately ran away, there were witnesses who saw the assaults, the person who was assaulted did not see the perpetrator, and the assaults caused minor injuries. The cases involved different perpetrators and victims (i.e., the cases were unrelated).

We manipulated how many people witnessed the crimes: In one case, two people saw the assault; and in the other case six people saw the assault. We counterbalanced which case had two, and which case had six witnesses. To justify the different number of witnesses, we changed the day and time in the scenarios. Subjects who were told there were six witnesses to the bar assault were told that the assault occurred on a Friday night when the bar was busy. In contrast, subjects who were told there were two witnesses who saw the bar assault were told that the assault occurred on a Monday afternoon when the bar was not busy.

After reading each case, subjects read the descriptions of the perpetrators, which, like Experiment 3a and 3b, included height, age, hair, and ethnicity. After reading each description, subjects read that the police had identified a man who lived close to the bar (in Case 1) and a man who worked close to the parking lot (in Case 2) as suspects, and considered them their prime suspects. Police then created two lineups using a photo of each suspect and seven other similar looking people. Subjects viewed the lineups and read that each eyewitness viewed the lineup, independently from each other, and were asked if the

perpetrator was present. In both cases, all of the witnesses had identified the same person from the lineups—in both cases, the suspect that had been identified by the police.

Subjects then read that afterwards, in a periodic review of police procedures, the officer investigating both cases had not followed the police guidelines for administering lineups. Throughout his career, the officer told witnesses—prior to them making a decision—which person in the lineup was the police suspect. We also told subjects that "the police guidelines state police should not tell witnesses who the suspect is at any stage in the lineup procedure".

Subjects then indicated the probability the police suspect was guilty, using a sliding scale (from 0-100), and by giving a verdict (guilty or not guilty).

Subjects answered the same questions we asked in the previous experiments, to determine if they could recall the number of witnesses and the questions to identify those who failed to comply with our experimental instructions.

Results.

***Compliance checks*.** A total of 20.6% did not comply with experimental instructions. These responses did not change the overall patterns in subsequent analyses, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question. See Appendix C for the subsequent analyses excluding those subjects who failed the compliance questions.

Attention checks. We also analysed subjects who failed to recall the number of witnesses in each case: 65.2% of subjects failed at least one of these two questions. This percentage was much higher than our other experiments (see Table B2). Excluding these

responses did change the significance of some of our inferential statistics, which we display in Appendix C.

Primary analysis. We then addressed our first research question: To what extent do subjects adjust their first guilt ratings for one crime to determine their guilt ratings for another crime with greater or fewer witnesses? To answer this question, we ran a 2 x 2 mixed ANOVA, with order (2 then 6 or 6 then 2) and guilt rating (Case 1 rating and Case 2 rating) as independent variables. The means and standard deviations for each condition are presented in Table D4. We first tested the Order x Guilt Rating interaction, which was significant, $F(1, 202) = 7.52$, $p = .01$, Partial $\eta^2 = .04$. To further interpret this significant interaction, we ran two paired-samples *t*-tests. We examined the first and second guilt ratings for subjects in the 2 then 6 condition. The *t*-test was not significant, $t(105) = 1.48$, $p = .14$, Cohen's $d = 0.09$. Subjects gave similar ratings in the first case with two witnesses ($M = 43.08$, $SD = 24.61$) compared to the second case with six witnesses ($M = 45.33$, $SD = 27.41$). We then examined first and second guilt ratings for those subjects in the 6 then 2 condition. The *t*-test was significant, $t(97) = 2.31$, $p = .02$, Cohen's $d = 0.19$. Subjects gave higher guilt ratings in the first case with six witnesses ($M = 44.51$, $SD = 22.42$) than the second case with two witnesses ($M = 40.26$, $SD = 21.23$). However, despite the statistical significance, the difference between the means in the guilt ratings between cases was small, and indeed is categorised as a small effect size according to Cohen (1988).

Although it was not our research question, we wondered if the order in which subjects were presented the two cases with differing number of witnesses influenced their guilt ratings. To answer this question, we tested the main effect of order, which was not significant, $F(1, 202) = .33$, $p = .57$, Partial $\eta^2 = .002$. As Figure 5 suggests, there was no

evidence that subjects who read about a case with six witnesses first and then a case with two witnesses second gave significantly different guilt ratings compared to those subjects who read about a case with two witnesses first and then a case with six witnesses second. This finding is consistent with Experiment 1, in which we found no evidence of order effects.

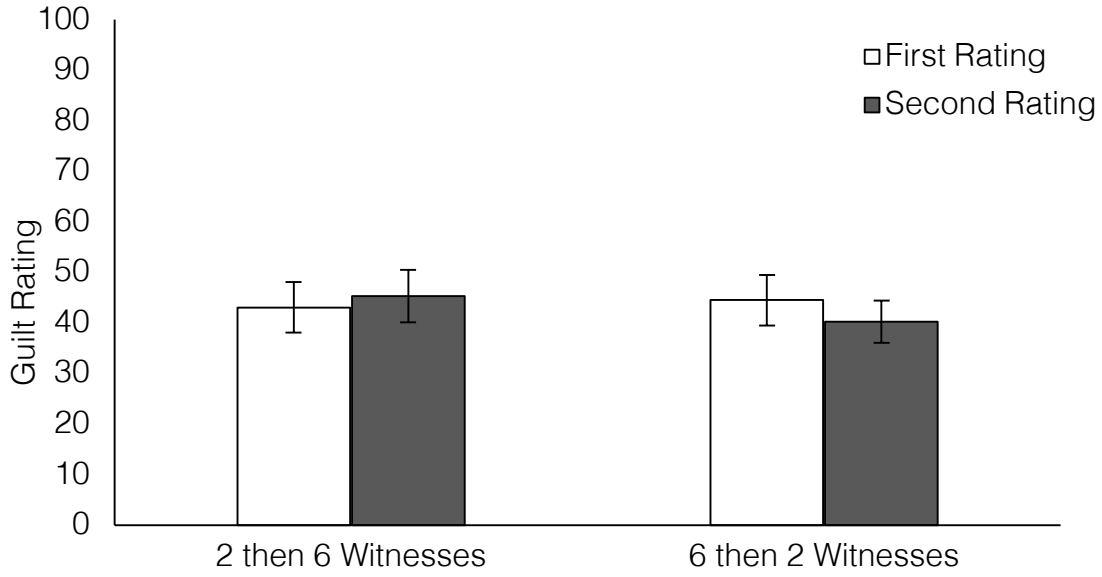

Figure 5. Mean guilt ratings for each witness condition in Experiment 4. The light bars represent subjects' Case 1 ratings. The dark bars represent subjects' Case 2 ratings. Error bars represent 95% confidence intervals of the cell means.

We examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes. We included reasons mentioned by at least 5% of subjects. The four reasons mentioned by at least 5% of subjects were the same for both Case 1 (Table 5a) and Case 2 (Table 5b). Interestingly, the percentage of subjects in the two and six witness conditions that mentioned the number of witnesses was similar for their Case 1 explanations ($Z = 0.19$, $p = .85$). There were no significant shifts in percentages of subjects mentioning the number of witnesses in either condition.

Table 5a

Reasons for Case 1 Guilt ratings in Experiment 4

	All Subjects	2 witnesses	6 witnesses
The police officer biased subjects' responses	55.4	59.4	51.0
Because witnesses identified the person	17.6	15.1	20.4
Uncertainty or there was not enough evidence presented	12.7	13.2	12.2
Mention the number of witnesses who identified the person	9.8	9.4	10.2

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects for Case 1 are listed. The "2 witnesses" condition presented here becomes the "2 then 6" condition in Table 5b after the second case is presented. Similarly, the "6 witnesses" condition here becomes the "6 then 2" condition in Table 5b.

Table 5b

Reasons for Case 2 Guilt ratings in Experiment 4

	All Subjects	2 then 6	6 then 2
The police officer biased subjects' responses	50.0	49.1	47.2
Because witnesses identified the person	18.1	23.6	11.3
Uncertainty or there was not enough evidence presented	14.7	12.3	16.0
Mention the number of witnesses who identified the person	10.3	14.2	5.7

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects for Case 2 are listed.

Discussion.

We found that subjects who were presented with a case involving two witnesses identifying the suspect did not subsequently give higher guilt ratings when presented with a case involving six witnesses identifying the suspect. This is not consistent with neither the

anchoring and adjustment nor repetition explanations. However, subjects who were presented with a case involving six witnesses identifying the suspect did give significantly lower guilt ratings when presented with a case involving two witnesses. This finding was consistent with the anchoring and adjustment explanation. The validity of these findings is, however, questionable. We unexpectedly observed that 65.2% of subjects failed to recall the number of witnesses in each of the two cases. This failure rate was almost double the failure rate in Experiments 1 and 2. Excluding these subjects did change our main findings—however these patterns are unlikely to be reliable too because the excluded subjects left approximately 35 subjects per condition. It is likely the reason why a large number of subjects failed the attention checks is because they confused the number of witnesses associated with each case. That is, it may be that our methodology was especially confusing for the subjects to process.

Another issue we recognise with this experiment is the potential confound of our justification for witness numbers with the differences in the number of witnesses. Recall, for the first case in the bar, we told subjects in the condition with six witnesses that it was a Friday night and the bar was busy. In the condition with two witnesses, we told subjects that it was a Monday afternoon and the bar was not busy. In hindsight, there was a clear issue with these justifications because they confounded with the number of witnesses. It is plausible that subjects might have thought witnesses at a busy bar on a Friday night would likely be judged as less reliable than the witnesses at a bar on a Monday afternoon due to alcohol consumption and how busy the bar was at the time of the assault. We did find some subjects explicitly mentioned the witnesses at the bar might have been impaired from alcohol consumption. Therefore, subjects in the 6 then 2 condition might have contrasting

six intoxicated witnesses with two sober witnesses, in which the pattern observed on Figure 5 is likely to be smaller than the difference between six sober witnesses and two sober witnesses. That is, if the time and day were removed as confounding variables, we might have seen a bigger downward adjustment in the 6 then 2 condition.

It is also worth discussing the effects of repetition. In this experiment, we removed the influence of repeatedly identifying the same suspect when additional witnesses came forward (unlike Experiments 1 and 2). However, witness identification evidence was still repeated twice within the experiment—once for Case 1 and once for Case 2. So, although we might have eliminated the effects of repeatedly identifying the suspect within a single case, it could be that generally mentioning witness evidence twice makes that particular type of evidence more salient and accessible to subjects, which in turn would influence their guilt ratings (Foster, Garry, Loftus, 2012; Alter & Oppenheimer, 2009; Bacon, 1979). If that were true, in this experiment we would expect that repetition would have influenced both guilt ratings because both case details were presented before making either of the guilt ratings. However, we have no way of testing this effect because we did not have a comparable control condition in which only one piece of witness evidence was presented. Indeed, it would have been impossible to create this control condition without introducing other confounding variables.

Taken together, Experiment 4 was largely unsuccessful at addressing our research questions because of the high number of people who failed to remember how many witnesses were associated with each of the two cases. However, we though the idea of comparing increases of additional witnesses with decreases in witnesses was promising because it allowed us to determine if subjects equally adjusted their guilt ratings in either

direction. According to the story model of juror decision-making and the research on confirmation bias, if subject's initial belief after reviewing the evidence is that the suspect is guilty, then they are likely to rely on, and put more weight on, evidence that confirms that belief than evidence that does not confirm that belief (Carlson & Russo, 2001; Russo, 2014; Russo, Medvec & Meloy, 1996). Experiments 1 – 3 support this explanation: Guilt ratings have been generally high on the 100-point scale (with guilt rating cell means ranging from 54.69 and 89.42, indicating a general belief that the suspect is guilty. If that is the case, we might expect subjects to be reluctant to decrease their guilt ratings when faced with evidence that does not confirm their belief of guilt. This finding could be experimentally tested by both increasing and decreasing the number of witnesses by the same magnitude and observing the shifts in guilt ratings. If subjects are more receptive to guilt-confirming information, then we would expect those subjects who were told that the number of witnesses increased would increase their guilt ratings to a greater extent than those subjects who were told that the number of witnesses decreased. In Experiment 5 we tested that idea.

Experiment 5

In Experiment 4 we found limited evidence that people use their first guilt ratings (based on one crime) to inform their second guilt ratings (based on a different, unrelated, crime). We also found an unusually high number of people who could not recall how many witnesses identified the suspect. Perhaps subjects found remembering information from two cases difficult. We wanted to test the extent to which subjects adjusted upwards and downwards from their first guilt rating at Time 1, when evaluating the same crime—not two different crimes. In Experiment 5, subjects were presented with a crime and gave guilt

ratings, then subjects were told that there had been an administrative error in the case details, and that the number of witnesses was either four more or four less than they were initially told.

Method.

Subjects. A total of 205 subjects completed the experiment using Amazon Mechanical Turk (www.mturk.com). Of these 205 subjects, 47.8% identified as women; 52.2% identified as men. Subjects were between 19 and 74 years of age (M_{age} = 35.40, SD_{age} = 11.46, $Median_{age}$ = 32.00), and received 1.00 USD upon completion.

Design. We used a 2 x 2 mixed design, with number of witnesses (10 then 6 or 2 then 6) and guilt rating (Time 1 and Time 2) as independent variables, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure.

Procedure. Subjects read the same instructions as Experiment 4 and viewed the bank robbery crime and (unbiased) lineup—the same crime as subjects were presented in Experiment 3. We manipulated the number of eyewitnesses; either two or ten. Like the previous experiments, all witnesses identified the police suspect from the lineup. Subjects then made the same guilt ratings and attention check questions as Experiment 4. Both groups then read that there had been an administrative error on the case file given to them, and that there were actually six witnesses (not the two or ten they were initially told). It was confirmed that these six witnesses unanimously and independently identified the same suspect. Subjects then provided guilt ratings and the attention check questions based on the correct information.

Results.

Compliance checks. We first addressed compliance: 30.2% did not comply with experimental instructions. These responses did not change the overall patterns in subsequent analyses, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question. See Appendix C for the subsequent analyses excluding those subjects who failed the compliance questions.

Attention checks. A total of 33.2% of subjects failed to recall at least one of our four manipulation questions. This percentage was closer to failure rates in our other experiments (see Table B2) than the failure rate in Experiment 4, suggesting our manipulation in the current experiment was easier for subjects to remember. Responses from these subjects who failed to recall at least one of these questions did not change the overall patterns in subsequent analyses, so we retained them in the dataset. The subsequent analyses excluding these subjects are also presented in Appendix C.

Lineup fairness. Lineup and lineup administration fairness ratings were analysed. Subjects in the 10 then 6 condition gave lower fairness ratings (M = 5.76, SD = 1.30) and lower administration fairness ratings (M = 5.54, SD = 1.40) than subjects in the 2 then 6 condition (M = 6.10, SD = 1.08; M = 5.88, SD = 1.15, respectively); $F(1, 205) = 4.23$, p = .04, Cohen's d = 0.28, $F(1, 205) = 3.60$, p = .06, Cohen's d = 0.27. Although, we did not intend to manipulate lineup or administration fairness, these findings make sense because subjects in the condition in which witnesses are removed might be generally more sceptical about the investigation and more likely assume the lineup was biased.

Number of witnesses between-subjects. Finally, before addressing our main research question, we tested the effect of the number of witnesses on guilt ratings by

comparing Time 1 guilt ratings of both conditions. Recall, in Experiment 1, using a biased lineup, we found no evidence that subjects who were told ten witnesses identified the suspect gave higher guilt ratings than subjects who were told two witnesses identified the suspect. Similarly, in Experiment 3, using both a biased and non-biased lineup, we found no evidence that people who were told twenty witnesses identified the suspect gave higher guilt ratings than people who were told fifteen, ten, or five witnesses identified the suspect. But, inconsistent with these previous experiments, and as Figure 6 shows, there was a statistically significant difference between Time 1 guilt ratings for the two conditions, which was confirmed with an independent samples *t*-test, $t(203) = 5.47$, $p < .001$, Cohen's $d = 0.77$. Subjects who were told ten witnesses identified the suspect gave higher guilt ratings ($M = 77.64$, $SD = 19.16$) than subjects who were told two witnesses identified the suspect ($M = 61.81$, $SD = 22.06$). This finding is consistent with previous research manipulating the number of plaintiffs in a case (Horowitz & Bordens, 1988; 2000) But, this finding is not consistent with Experiment 1 and Experiment 3.

***Within-subjects guilt rating adjustments*.** We then addressed our main research question: To what extent did subjects adjust upwards or downwards from their Time 1 guilt rating after learning about the administration error? To answer this question, we ran a 2 x 2 mixed ANOVA, with number of witnesses (10 then 6 or 2 then 6) and guilt rating (Time 1 and Time 2) as independent variables, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure. The means and standard deviations for each condition are presented in Table D5. We tested the Guilt Rating x Number of Witnesses interaction, which was significant, $F(1, 203) = 72.82$, $p < .001$, Partial $\eta^2 = .26$. To further interpret this interaction, we ran two paired-samples *t*-tests. We

examined the Time 1 and Time 2 guilt ratings for the subjects in the 2 then 6 condition. The *t*-test was significant, $t(105) = 8.66$, $p < .001$, Cohen's $d = 0.98$, which is consistent with Experiments 1 and 2, showing that subjects who were told two witnesses identified the suspect ($M = 61.81$, $SD = 22.06$) significantly adjusted their guilt ratings upwards after being told there was an error and there were actually six witnesses ($M = 83.24$, $SD = 21.76$). Second, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 10 then 6 condition. The *t*-test was significant, $t(98) = 2.29$, $p = .024$, Cohen's $d = 0.16$. Subjects who were told ten witnesses identified the suspect ($M = 77.64$, $SD = 19.16$) significantly adjusted their guilt ratings downwards after being told there was an error and there were actually six witnesses who identified the suspect ($M = 74.56$, $SD = 20.09$). This finding is consistent with our explanation that subjects used their first ratings as an anchor and adjusted their subsequent guilt ratings.

Although we found evidence that subjects adjusted their guilt ratings both upwards and downwards, Figure 6 suggests that these adjustments were not equivalent. Specifically, subjects in the 2 then 6 condition adjusted upwards, demonstrated by large effect size, (Cohen's $d = 0.98$), to a greater extent than subjects in the 10 then 6 adjusted downwards, demonstrated by a much smaller effect size (Cohen's $d = 0.16$). We also confirmed this finding by running an independent samples *t*-test with number of witnesses as the independent variable and Time 2 guilt rating as the dependent variable, $t(203) = 2.96$, $p = .003$, Cohen's $d = 0.41$. Guilt ratings were higher for subjects when they were told there were six witnesses for those in the 2 then 6 condition ($M = 83.24$, $SD = 21.76$) than the 10 then 6 condition ($M = 74.56$, $SD = 20.09$). That is, even though both groups were told six witnesses unanimously identified the suspect, the guilt ratings were different depending on

whether they were initially told a greater number of witnesses identified the suspect or fewer witnesses identified the suspect.

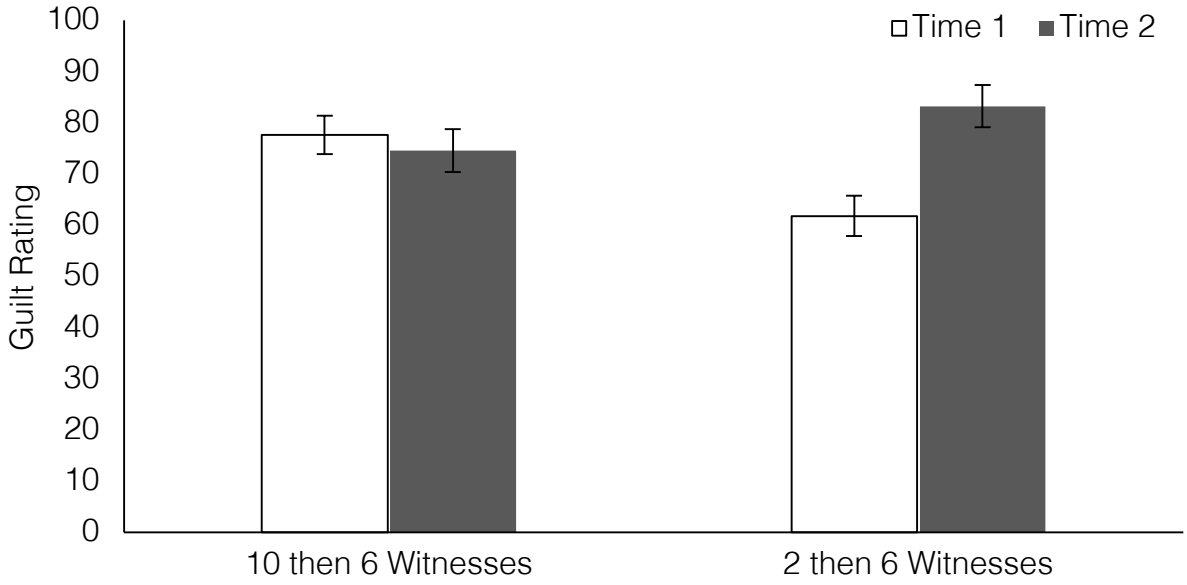

Figure 6. Mean guilt ratings in each witness condition in Experiment 5. The light bars represent subjects' Time 1 guilt ratings. The dark bars represent subjects' Time 2 guilt ratings. Error bars represent 95% confidence intervals of the cell means.

We examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes. We displayed the reasons mentioned by at least 5% of subjects by reasons for the guilt ratings associated with Time 1 and Time 2, on Table 6a and Table 6b, respectively. Subjects who were told there were ten witnesses at Time 1 were more likely to mention the witness evidence (71.7%) than subjects who were told there were two witnesses at Time 1 (50.0%) ($Z = 3.18$, $p = .001$). When the number of witnesses then decreased in the 10 then 6 condition, subjects were less likely to mention the witness evidence at Time 2 (56.6%) than they were at Time 1 (71.7%) ($Z = 2.22$, $p = .03$). Conversely, when the number of witnesses then increased in the 2 then 6 condition, subjects were more likely to mention the witness evidence at Time 2 (65.1%)

than they were at Time 1 (50.0%) ($Z = 2.22$, $p = .03$). The percentage of subjects mentioning the specific number of witnesses who identified the suspect was similar at Time 1 between the 2 then 6 and 10 then 6 conditions (44.3% and 43.4% respectively, $Z = 0.13$, $p = .90$). At Time 2, the percentage of subjects mentioning the specific number of witnesses who identified the suspect then shifted depending on the condition. At Time 2, more subjects in the 2 then 6 mentioned the specific number of witnesses than in the 10 then 6 condition ($Z = 3.55$, $p < .05$). We found that 15.1% of subjects mentioned at Time 2 that the suspect's guilt was positively associated with the likelihood of that person being guilty. This percentage was similar in both the 2 then 6 and 10 then 6 conditions ($Z = 0.20$, $p = .84$).

Table 6a

Reasons for Time 1 Guilt ratings in Experiment 5

Reasons	All subjects	2 witnesses	10 witnesses
Because witnesses identified the person	62.4	50.0	71.7
Mention the number of witnesses who identified the person	45.4	44.3	43.4
Because the witnesses were unanimous	24.4	2.8	45.5
Uncertainty or there was not enough evidence presented	14.6	17.0	11.1
Eyewitness evidence or eyewitness memory is unreliable	14.1	17.0	10.1
Multiple people in the lineup matched the description	8.8	13.2	1.0
The person the witnesses identified matched the description of the perpetrator	7.3	8.5	5.1

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 1 are listed. The "2 witnesses" condition presented here becomes the "2 then 6" condition in Table 6b after additional witnesses are presented. Similarly, the "6 witnesses" condition here becomes the "6 then 2" condition in Table 6b.

Table 6b

Reasons for Time 2 Guilt ratings in Experiment 5

	All subjects	2 then 6	10 then 6
Because witnesses identified the person	61.5	65.1	56.6
Mention the number of witnesses who identified the person	42.4	53.8	29.3
Because the witnesses were unanimous	24.4	18.9	29.3
Positive relationship between number of witnesses and guilt	15.1	15.1	14.1
Uncertainty or there was not enough evidence presented	8.3	7.5	9.1

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 2 are listed.

Discussion.

In Experiment 5, we found that subjects adjusted their guilt ratings, consistent with the anchoring and adjustment explanation. We also found differences in initial guilt ratings when there were ten compared to two witnesses. Recall, in Experiment 1, we saw no evidence of differences between two and ten witnesses identifying the suspect. Of course, Experiment 1 differed in methodology to the current experiment. In particular, we used different crimes and the lineups in Experiment 1 were biased, but were not biased in the current experiment. Therefore, it could be that the function between the number of

witnesses and guilt ratings differs based on the extent to which the lineup is biased. That is, the lack of difference in Time 1 guilt ratings in Experiment 1 was due to the lineup bias. However, we partly addressed this concern in Experiment 3, where we tested biased and non-biased lineups over five, ten, fifeteen, and twenty witnesses. But Experiment 3 did not specifically compare two witnesses to ten witnesses. It could be that the number of witnesses does influence initial guilt ratings for non-biased lineups—but the function is non-linear. Specifically, the effect of more witnesses identifying the suspect has a diminishing effect on subject's guilt ratings. In Experiment 6, we tested this explanation.

In Experiment 5, we also found people increased their guilt ratings to a greater extent after the number of witnesses increased than the extent to which subjects decreased their guilt ratings after the number of witnesses decreased. These findings are not necessarily inconsistent with our anchoring adjustment explanation. However, anchoring and adjustment does not explain why we observed asymmetrical adjustments based on whether witnesses are added or removed during the experiment.

We propose three explanations to help explain this asymmetrical effect observed in Experiment 5. I have outlined the first explanation in previous experiments: Repeatedly identifying the suspect might make that witness evidence more salient to subjects, which might make subjects increase their guilt ratings (Zaragoza & Mitchell, 1996; Wright, Wade, Watson, 1996; Wang, Brashier, Wing, Marsh, & Cabeza, 2016).

The second explanation is that after subjects are told that witnesses identified the suspect, that evidence makes them believe the suspect is probably guilty. When subjects are presented evidence that confirms this initial belief, they use that information and update their guilt ratings. But when subjects are presented with evidence that disconfirms this

initial belief, they are more likely to disregard the new evidence and do not update their guilt ratings to the same extent. This selective use of evidence is known as a confirmation bias—a persuasive bias that affects many decisions people make (Nickerson, 1998). This explanation is also consistent with the story model of juror decision-making (Pennington & Hastie, 1981, 1986, 1988, 1993).

The third explanation for why we observed the asymmetrical pattern of findings on Figure 6 is a similar explanation to the possible explanation for difference in initial guilt ratings that I described earlier. Perhaps the influence of the number of witnesses on guilt ratings is not a linear relationship. That is, the impact of additional witnesses coming forward has a decreasing effect on guilt ratings. It could be that the function of guilt ratings starts to plateau after only a small number of agreeing witnesses, perhaps two or three witnesses. This explanation is consistent with the conformity literature. In Asch's conformity experiments (1951; 1955; 1956), he found that increased group size increased conformity when he compared the influence of one, two, and three members, but group sizes greater than three did not continue to increase subject conformity. This non-linear function is also consistent with Gunn et al. (2016), who found non-linear diagnosticity functions for the number of unanimous witnesses identifying a single suspect from a biased lineup. Perhaps the influence from witnesses on subjects in our experiment did not continue to increase after three witnesses. Therefore, the guilt ratings in the 10 then 6 condition would be similar because both would correspond to flatter parts of the function between number of witnesses and guilt ratings. In contrast, guilt ratings in the 2 then 6 condition would be different because the corresponding function at these lower numbers should be steeper.

A non-linear relationship between the number of witnesses and guilt ratings for non-biased lineup might help explain our two major findings in Experiment 5. Namely, the difference in initial guilt ratings and the asymmetrical shifts in guilt ratings once the number of witnesses changed. Therefore, it was this explanation that we tested in Experiment 6.

Experiment 6a and 6b

In Experiment 6 we addressed this third explanation that the number of witnesses has a diminishing effect on guilt ratings— that the effect of adding additional witnesses plateaus after two witnesses—by running Experiment 3a and 3b again, but instead of using five, ten, fifteen, and twenty witnesses, we used one, two, three, and four witnesses. Because Experiment 6a and 6b were replications of Experiment 3a and 3b, we reintroduced the bias variable into this experiment. We could therefore determine if the number of witnesses and lineup bias affected guilt ratings differently over a greater range of witness numbers. If the effect of witness numbers on guilt ratings does plateau at two witnesses, then the two condition should have a higher guilt rating than the one condition, and the three condition should have a higher guilt rating than the two condition, but the four condition and the three condition should have similar guilt rating means. Moreover, the mean guilt ratings for the three condition and four condition should be similar (and non-significantly different) from the five, ten, fifteen, and twenty condition in Experiment 3a and 3b. In other words, we think the number or witnesses has a non-linear effect on guilt ratings, which plateaus after approximately two or three witnesses identify the suspect.

Method.

Subjects. A total of 207 subjects in Experiment 6a and 205 in Experiment 6b completed the experiment using Amazon Mechanical Turk (www.mturk.com). Of these subjects, 39.6% in Experiment 6a and 40.0% in Experiment 6b identified as women; 60.4% in Experiment 6a and 59.5% in Experiment 6b identified as men. Subjects were between 19 and 71 years of age in Experiment 6a and 20 and 66 in Experiment 6b (Experiment 6a: M_{age} = 35.19, SD_{age} = 10.12, $Median_{age}$ = 32.00. Experiment 6b: M_{age} = 37.32, SD_{age} = 10.66, $Median_{age}$ = 34.00). Subjects received 1.00 USD upon completion.

Design. We used a 4 x 2 design, with number of witnesses (1, 2, 3, or 4) and lineup administration (biased or not biased) as independent variables, with the lineup administration variable manipulated between Experiment 6a and 6b. Guilt rating was the dependent variable.

Procedure. In the interest of open science practices, we pre-registered our hypotheses, analyses, sample size, and exclusion criteria on the online repository AsPredicted (aspredicted.org), see Appendix E. Experiment 6a was identical to Experiment 3a and Experiment 6b was identical to Experiment 3b, with the following change: the number of witnesses who saw the crime and identified the suspect in the current experiment were one, two, three, or four.

Results.

Compliance checks. We first addressed compliance: 41.0% (49.3% in Experiment 6a and 32.7% in Experiment 6b) did not comply with experimental instructions. These responses did not change the overall patterns in subsequent analyses, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of

subjects who failed each question. See Appendix C for the subsequent analyses excluding those subjects who failed the compliance questions.

Attention checks. A total of 38.3% (48.3% in Experiment 6a and 28.3% in Experiment 6b) of subjects failed to recall at least one of our four questions about the number of witnesses who identified the suspect. As Table B2 shows, these percentages were higher than those in Experiment 3, which we think might be caused by a higher number of nonsense responses in the current experiment. At the time we ran the current experiment, many researchers were claiming that bots were completing studies on Mechanical Turk and giving nonsense answers (Dreyfuss, 2018). Despite the higher percentages, responses from these subjects who failed to recall at least one of these questions did not change the overall pattern of results, so we retained them in the dataset. The subsequent analyses excluding these subjects are also presented in Appendix C.

Lineup fairness. Lineup and lineup administration fairness ratings were analysed. Subjects who were told the lineup administration was biased gave lower lineup fairness ratings ($M = 4.94$, $SD = 2.38$) and lower administration fairness ratings ($M = 4.68$, $SD = 2.53$) than those subjects who were not told the lineup administration was biased (lineup fairness: $M = 5.71$, $SD = 1.32$) (lineup administration fairness: $M = 5.70$, $SD = 1.30$), lineup fairness: $F(1, 416) = 24.03$, $p < .001$, Partial $\eta^2 = .06$; lineup administration fairness: $F(1, 411) = 36.77$, $p < .001$, Partial $\eta^2 = .08$. There were no other significant main effects or interactions (all $p > .18$). All of these non-significant main effects and interactions are presented in Appendix C.

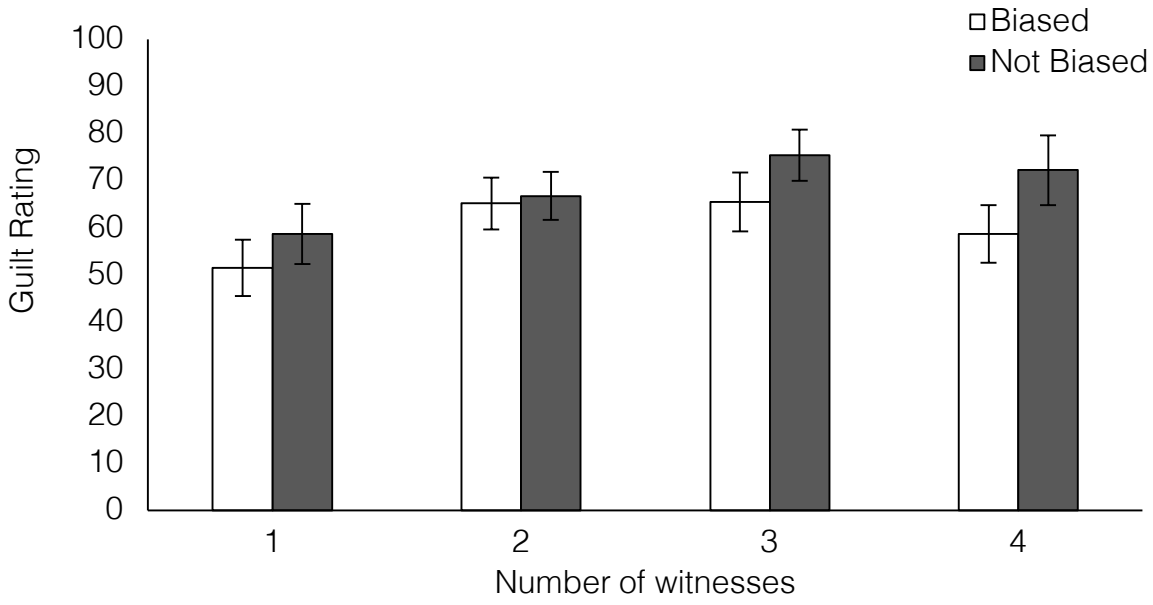

Figure 7. Mean guilt ratings for each witness condition in Experiment 6. The light bars represent subjects in Experiment 6a, who were told a police officer identified the suspect to witnesses. The dark bars represent subjects in Experiment 6b, who were not given any information about how the police officer administered the lineup. Error bars represent 95% confidence intervals of the cell means.

Primary analysis. We then addressed our primary research question: To what extent does the number of witnesses influence guilt ratings when the number of witnesses was small? To answer this question, we ran a 4 x 2 ANOVA, with number of witnesses (1, 2, 3, or 4) and lineup administration (biased or not biased) as independent variables, with guilt ratings as the dependent variable. The means and standard deviations for each condition are presented in Table D6. First, the Number of Witnesses x Lineup Administration interaction was non-significant, $F(3, 404) = 1.31$, $p = .27$, Partial $\eta^2 = .01$. Second, and as the Figure 7 shows, the main effect of number of witnesses was significant, $F(3, 404) = 9.04$, $p < .001$, Partial $\eta^2 = .06$. To further examine this main effect, we ran

Tukey post-hoc tests. Subjects who were told one witness identified the suspect gave lower guilt ratings (M = 54.91, SD = 23.13) than those subjects who were told two witnesses (M = 65.97, SD = 19.04) (p = .002), three witnesses (M = 70.35, SD = 21.92) (p < .001), and four witnesses (M = 65.27, SD = 25.36) (p = .004) identified the suspect. There were no other significant differences. That is, after two witnesses, there was no evidence that more witnesses increased subjects' guilt ratings.

Bias. We then addressed our second research question: To what extent did biased and non-biased lineups influence subjects' guilt ratings? To answer this question, we used the 4 x 2 ANOVA reported above. Recall, the Number of Witnesses x Lineup Administration interaction was non-significant, suggesting that the lineup administration does not lead to a different pattern of responding over the differing numbers of witnesses. But we found a main effect of lineup, $F(1, 404) = 13.58$, $p < .001$, Partial $\eta^2 = .03$. As Figure 7 shows, those in Experiment 6a—the biased administration—gave lower guilt ratings (M = 59.95, SD = 22.52) than subjects in Experiment 6b—the non-biased administration (M = 68.18 SD = 23.06). This finding was consistent with Experiment 3.

Comparison to Experiment 3. To compare the effects of number of witnesses and bias across a larger range of groups sizes, we combined the data from Experiments 3 and 6—which were identical except for varying group size. We ran an 8 x 2 ANOVA, with number of witnesses (1, 2, 3, 4, 5, 10, 15, or 20) and lineup administration (biased or not biased) as independent variables, with guilt rating as the dependent variable. There was a significant interaction between the number of witnesses and whether the lineup administration was biased, $F(7, 878) = 2.31$, $p = .03$, Partial $\eta^2 = .02$. As Figure 8 shows, in the non-biased conditions, the number of witnesses influenced guilt ratings. In the non-

biased conditions, subjects who were told there were three, four, five, ten, fifteen, and twenty witnesses gave significantly higher guilt ratings than those subjects who were told there was one witness (all < .05). Those in the ten, fifteen, and twenty witness conditions also gave significantly higher guilt ratings than subjects who were told that there were two witnesses (all < .05). Although not all differences displayed in Figure 8 were significant, the pattern shows that the addition of witnesses influences guilt ratings the most when there are lower numbers of witnesses. As the number of witnesses increases, the influence of adding more witnesses has less of an effect on guilt ratings. In the biased conditions, there was no evidence that the number of witnesses influenced guilt ratings. There were no significant group differences in guilt ratings between all eight conditions (all *p*-values > .05). Recall, in both Experiment 3 and Experiment 6, this Administration Bias x Number of Witnesses interaction was not significant—only when we combined the data from both experiments, did we observe this significant interaction. Therefore, the number of witnesses seems to influence guilt ratings only when the lineup administration is not biased and number of total witnesses one or two.

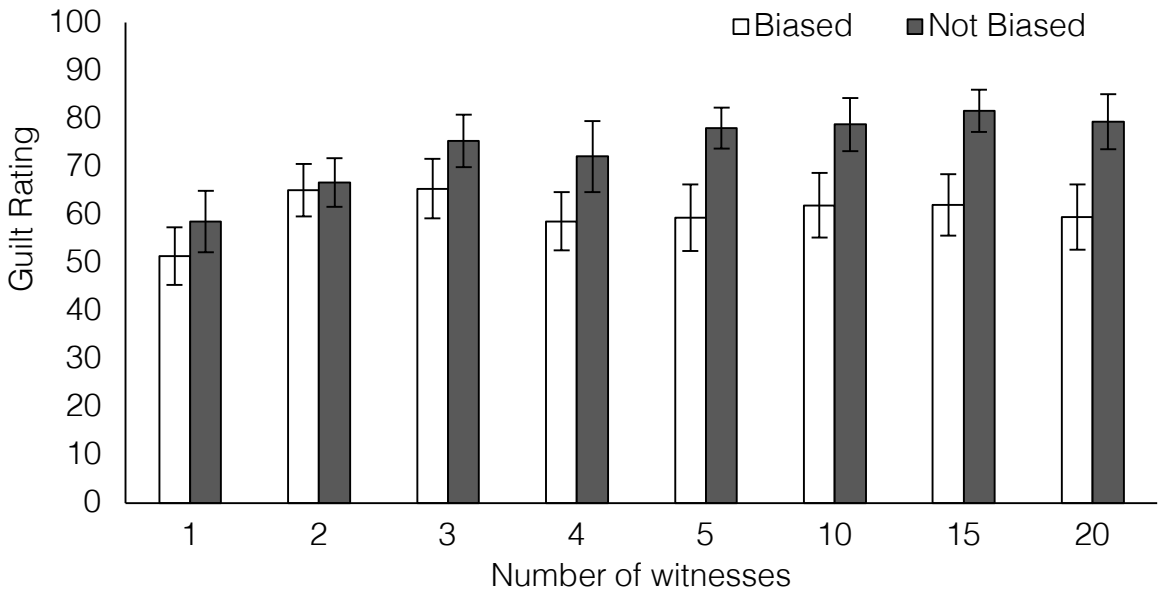

Figure 8. Mean guilt ratings for subjects in each witness condition for Experiments 3 (5, 10, 15, or 20 witnesses) and Experiment 6 (1, 2, 3, or 4 witnesses). The light bars represent subjects in Experiments 3a or 6a, who were told a police officer identified the suspect to witnesses. The dark bars represent subjects in Experiments 3b or 6b, who were not given any information about how the police officer administered the lineup. Error bars represent 95% confidence intervals of the cell means.

One possible reason for the significant interaction between bias and number of witnesses is that people in the biased condition are more uncertain about guilt, and represent this uncertainty by giving a guilt rating of 50% probability. Indeed, some research shows that people use 50% on scales to represent the verbal expression "a 50/50 chance" rather than actually meaning 50% probability (de Bruin, Fischhoff, Millstein, & Halpern-Felsher, 2000). The combined data from Experiment 3 and 6 support this explanation: More subjects in the biased administration conditions gave 50% guilt ratings (10.5%) compared to subjects in the not biased administration conditions (4.0%) (Z =

3.80, $p < .001$). Furthermore, excluding these subjects from the combined ANOVA eliminates the significant Lineup Administration x Number of Witnesses interaction, $F(7, 808) = 1.70$, $p = .11$, Partial $\eta^2 = .02$. Despite these findings, we cannot conclude that all of those subjects who gave a 50% guilt ratings used it to express uncertainty—some subjects might have though there was actually a 50% probability the suspect was guilty.

We examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes, and displaying the reasons mentioned by at least 5% of subjects on Table 7. Those subjects who were told two witnesses identified the suspect were more likely to mention the witness evidence than subjects who were told one witness identified the suspect (biased: $Z = 2.67$, $p = .01$; not biased: $Z = 4.74$, $p < .001$). The table also shows that none of the subjects who were told one witness identified the suspect mentioned the number of witnesses or unanimity. This makes sense because subjects often referred to the lone witness as "the witness" rather than specifically mentioning the number "one". Furthermore, it is obvious in the one witness condition that there are no other witnesses to be unanimous with. It is also noteworthy that the percentages of subjects mentioning the witness evidence increased steeply between the one and two witness conditions (Z tests presented above), but was similar in the three and four witness conditions (biased: $Z = 0$, $p = 1$ (identical means); not biased: $Z = .28$, $p = .78$). This pattern is consistent with the non-linear pattern observed in guilt ratings. The positive association between the number of witnesses identifying the suspect and subjects mentioning unanimity was only clearly present in the non-biased conditions, suggesting that perhaps some subjects realised that the unanimity might have been caused by the biased administration. The finding that no subjects in the non-biased conditions mentioned

a bias is consistent with Experiment 3 and further supports the idea that our lineup materials were not unintentionally biased. Finally, although in the previous paragraph we predicted that subjects in the biased conditions would be more uncertain of guilt than in the not biased conditions, this was not consistent with the percentage of subjects who explicitly mentioned uncertainty in their written descriptions.

Table 7

Reasons for Guilt ratings in Experiment 6

		Biased				Not Biased			
Reasons	All subjects	1	2	3	4	1	2	3	4
Because the witness(es) identified the person	61.2	1.8	16.7	19.2	19.2	13.7	58.5	64.7	62.0
Mention the number of witnesses who identified the person	33.8	0.0	20.8	11.5	9.6	2.0	54.7	51.0	44.0
Because the witnesses were unanimous	48.1	0.0	0.0	13.5	9.6	0.0	24.5	37.3	54.0
Uncertainty or there was not enough evidence presented	20.7	9.1	4.2	7.7	9.6	15.7	15.1	3.9	16.0
Police tainted/biased the evidence when they identified the suspect to the witness(es)	16.6	25.5	10.4	19.2	19.2	0.0	0.0	0.0	0.0
Multiple people in the lineup matched the description	8.1	10.9	2.1	3.8	3.8	17.6	5.7	7.8	14.0
Eyewitness evidence or eyewitness memory is unreliable	5.2	1.8	2.1	1.9	1.9	23.5	18.9	3.9	4.0
The person the witness(es) identified matched the description of the perpetrator	6.8	7.3	14.6	0.0	1.9	9.8	9.4	9.8	2.0

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects 1 are listed.

Discussion.

Experiment 6 showed the influence of the number of witnesses on guilt ratings. The effect of additional witnesses coming forward had a diminishing effect on guilt ratings. Subjects who were told that two witnesses identified the suspect gave significantly higher guilt ratings than subjects who were told that one witness identified the suspect. In contrast, subjects who were told that five witnesses identified the suspect did not give significantly higher guilt ratings than subjects who were that told four witnesses identified the suspect.

The results of this experiment can help us understand the discrepancy between the findings observed in Experiment 1 and Experiment 5. In both of these experiments we compared subjects who were told there were two witnesses with subjects who were told there were ten witnesses. Recall, in Experiment 1, we found no difference in these (Time 1) guilt ratings. But in Experiment 5, comparing the same numbers of witnesses, we observed higher guilt ratings for subjects in the ten witnesses group compared to the two witnesses group. The main difference between these two experiments was the fact that the lineups used in Experiment 1 were biased and the lineup used in Experiment 5 were not biased. The combined results of Experiment 3 and Experiment 6 show that the function of witness numbers and guilt ratings differ by whether the lineup was biased. Figure 8 shows a direct comparison between 2 and 10 witnesses—the white (biased) bars corresponding to the two and ten witness conditions are similar, but the dark (not biased) bar corresponding to the ten witnesses condition is significantly higher than the bar corresponding to the two witnesses condition. That is, the number of witnesses does seem

to have some between-subjects influence on guilt ratings for non-biased lineups, but not for biased lineups.

In the current experiment this diminishing influence of the number of witnesses on guilt ratings was observed between-subjects. But it is possible that we might see the same effect within-subjects when the number of witnesses changes during the experiment. Indeed, if that were the case, it could explain the asymmetrical shift in guilt ratings observed in Experiment 5. Recall, in that experiment, subjects increased their guilt ratings in the 2 then 6 condition. But subjects did not significantly decrease their guilt ratings in the 10 then 6 condition. This pattern could be explained—at least in part—by the non-linear function observed in Figure 8. If this explanation is true, then we might not observe the asymmetrical effect found in Experiment 5—or at least not to the same extent—if all of the witness groups sizes were greater than two. That is, if the number of witnesses is never below two—the point at which function seems to plateau—then any shifts in guilt ratings are less likely to be explained by the non-linear function, and more likely to be caused by the anchoring and adjustment and confirmation bias explanations.

Therefore, if we compare, for example, a 5 then 10 condition and a 15 then 10 condition, in the same manner as Experiment 5, we might observe symmetrical increases and decreases in guilt ratings for those respective conditions. However, there is reason to believe we might still observe an asymmetrical pattern with these higher numbers of witnesses. Recall, both the story model of juror decision-making and the confirmation bias predict that if subjects initially believe the suspect is guilty, they are more likely to attend and utilise evidence that supports that belief and disregard evidence that does not support that belief (Nickerson, 1998; Pennington & Hastie, 1981; 1986; 1988; 1993).

Specifically, if subjects believe the suspect is probably guilty, the addition of witness evidence that confirms that belief is likely to lead to increases in guilt ratings. In contrast, if subjects believe the suspect is probably guilty, the removal of witness evidence will be inconsistent with their belief—and subjects will be less likely to use that information to adjust their guilt ratings.

In Experiment 7 we tested this explanation by re-running Experiment 5, but comparing guilt ratings between a 5 then 10 and a 15 then 10 witness condition. If the non-linear function of guilt ratings observed in Experiment 6 extends to within-subjects differences, then using witness numbers of 5, 10, and 15 should largely remove this influence. We would therefore expect an elimination or a reduction in the asymmetrical effect observed in Experiment 5. If we observe a reduction in this asymmetrical pattern—and not the total elimination—then it is possible that our confirmation of guilt explanation is still influencing subjects. That is, subjects are more likely to use evidence that confirms guilt than evidence that does not confirm guilt.

Experiment 7

In Experiment 7, we ran the same procedure as Experiment 5, but we only used group sizes higher than two. Specifically, in one condition, subjects were told there were five witnesses, then they were subsequently told there were ten witnesses. In the other condition, subjects were told there were fifteen witnesses, then they were subsequently told there were ten witnesses. If the asymmetrical changes in guilt ratings observed in Experiment 5 were caused by the diminishing effect of the number of witnesses after a group size of two, then we should expect symmetrical differences in Experiment 7. That is, if five witnesses are added, it should have an equivalent but opposite effect on guilt ratings

as if five witnesses are removed. But if we observe a similar asymmetrical pattern after increasing witness numbers, it is likely to be caused by subjects being more influenced by evidence that confirms their belief that the suspect is guilty than evidence that does not confirm their belief.

Method.

Subjects. A total of 215 subjects completed the experiment using Amazon Mechanical Turk (www.mturk.com). Of these 215 subjects, 40.5% identified as women; 59.1% identified as men. Subjects were between 20 and 68 years of age (M_{age} = 36.28, SD_{age} = 10.53, $Median_{age}$ = 34.00), and received 1.00 USD upon completion.

Design. We used a 2 x 2 mixed design, with number of witnesses (15 then 10 or 5 then 10) and guilt rating (Time 1 and Time 2) as independent variables, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure.

Procedure. We pre-registered our hypotheses, analyses, sample size, and exclusion criteria on the online repository AsPredicted (aspredicted.org), see Appendix E. The procedure was identical to Experiment 5 except for the number of witnesses in each condition. Recall, in Experiment 5, there were two conditions: either two witnesses or ten witnesses. Then, subjects read that there was administration error and there were actually six witnesses. In Experiment 7, subjects read that either five witnesses or fifteen witnesses saw the crime. Then, all subjects read that there was administration error and there were actually ten witnesses.

Results.

Compliance checks. We first addressed compliance: 27.0% did not comply with experimental instructions. These responses did not change the overall patterns in subsequent analyses, so we retained them in the dataset. See Table B1 for a list of these compliance questions and the percentage of subjects who failed each question. See Appendix C for the subsequent analyses excluding those subjects who failed the compliance questions.

Attention checks. As Table B2 shows, 27.0% of subjects failed to recall at least one of our four manipulation questions. Responses from these subjects who failed to recall at least one of these questions did not change the overall pattern of results, so we retained them in the dataset. The subsequent analyses excluding these subjects are presented in Appendix C.

***Lineup fairness*.** Lineup and lineup administration fairness ratings were analysed. Subjects in the 15 then 10 condition gave similar fairness ratings (M = 6.04, SD = 1.23) and administration fairness ratings (M = 5.79, SD = 1.38) as subjects in the 5 then 10 condition (M = 6.26, SD = 1.00; M = 6.10, SD = 1.05, respectively), $F(1, 213) = 2.14$, p = .15, Partial η^2 = .20, $F(1, 213) = 3.41$, p = .07, Partial η^2 = .25. This finding is not consistent with Experiment 5. In both experiments, we did not manipulate lineup or administration bias. The small effect (Cohen's d = 0.27 and .28) of fairness ratings observed in Experiment 5 could be attributed to the general scepticism in the investigation when the number of witnesses were reduced. Although we did observe a similar pattern in the current experiment, it was not statistically reliable.

Number of witnesses between-subjects. Before addressing our main research question, we tested the effect of the number of witnesses on guilt ratings by comparing the

first guilt ratings of both conditions—when the 5 then 10 condition were told five witnesses identified the suspect and the 15 then 10 condition were told fifteen witnesses identified the suspect. Recall, in Experiment 3, using both a biased lineup and a non-biased lineup, we found no evidence subjects who were told fifteen witnesses identified the suspect gave higher guilt ratings than people who were told five witnesses identified the suspect. Consistent with Experiment 1 and 3, and as Figure 9 shows, there was no evidence of a difference between Time 1 guilt ratings for the two conditions, which was confirmed with an independent samples *t*-test, $t(213) = 1.44$, $p = .15$, Cohen's $d = 0.20$.

Primary analysis. We then addressed our main research question: To what extent do people adjust upwards or downwards from their initial guilt rating after learning about the administration error? To answer this question, we ran a 2 x 2 mixed ANOVA, with number of witnesses (15 then 10 or 5 then 10) and guilt rating (Time 1 and Time 2) as independent variables, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure. The means and standard deviations for each condition are presented in Table D7. We tested the Guilt Rating x Number of Witnesses interaction, which was significant, $F(1, 213) = 32.62$, $p < .001$, Partial $\eta^2 = .13$. To further interpret this interaction, we ran two paired-samples *t*-tests. First, we examined the first and second guilt ratings for the subjects in the 5 then 10 condition. The *t*-test was significant, $t(108) = 6.96$, $p < .001$, Cohen's $d = 0.52$. As Figure 9 shows, subjects who were first told five witnesses identified the suspect ($M = 77.21$, $SD = 20.59$) significantly adjusted their guilt ratings upwards after being told there was an error and there were actually ten witnesses who identified the suspect ($M = 87.38$, $SD = 18.15$). This finding is consistent with Experiments 1, 2, and 5, showing that increasing the number of witnesses,

within-subjects, increased guilt ratings. Second, we examined the Time 1 (M = 81.09, SD = 18.91) and Time 2 guilt ratings (M = 80.15, SD = 19.91) for the subjects in the 15 then 10 condition. The t-test was not significant, $t(105) = .74$, $p = .46$, Cohen's $d = 0.05$. This finding is not consistent with Experiment 5, in which subjects significantly adjusted downwards. However, the general asymmetrical pattern observed is similar to Experiment 5.

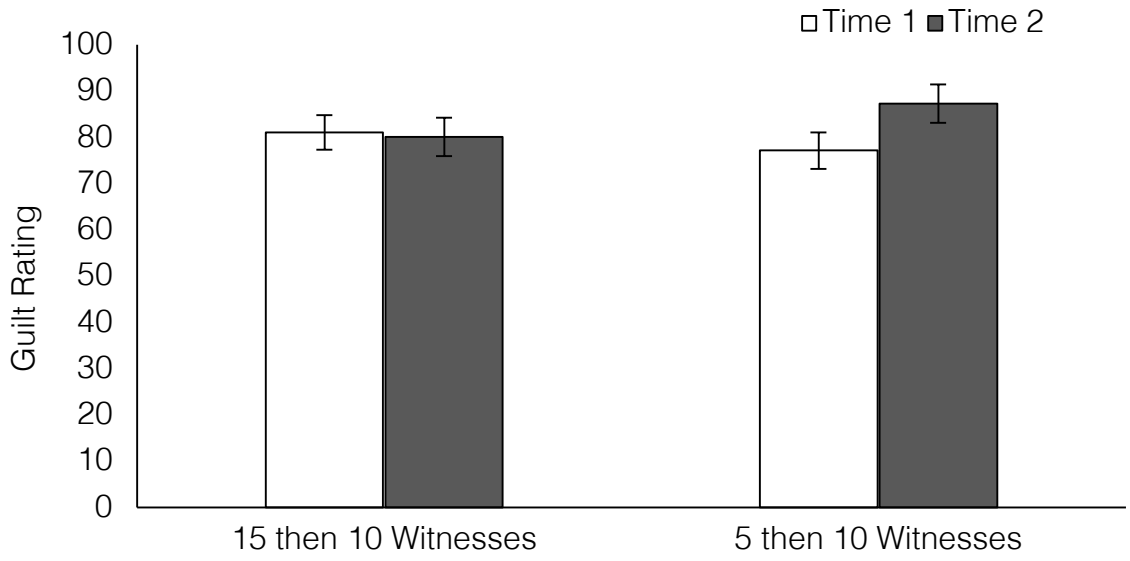

Figure 9. Mean guilt ratings for each witness condition in Experiment 7. The light bars represent subjects' Time 1 guilt ratings. The dark bars represent subjects' Time 2 guilt ratings. Error bars represent 95% confidence intervals of the cell means.

As in Experiment 5, we found that the extent to which subjects in the two conditions adjusted was significantly different. We confirmed this finding by running an independent samples t-test with number of witnesses as the independent variable and Time 2 guilt rating as the dependent variable, $t(213) = 2.78$, $p = .006$, Cohen's $d = 0.38$. Guilt ratings were higher for subjects when they were told there were ten witnesses for those in the 5 then 10 condition (M = 87.38, SD = 18.15) than the 15 then 10 condition (M = 80.15,

SD = 19.91). That is, even though both groups were told ten witnesses unanimously identified the suspect, the guilt ratings were higher if subjects were first told the number of witnesses was lower than 10 compared to first being told the number of witnesses was higher than 10.

We examined the explanations subjects gave for their guilt ratings by coding responses and grouping them into common themes. We displayed the reasons mentioned by at least 5% of subjects, and broken up by reasons for the guilt ratings associated with Time 1 and Time 2, on Table 8a and Table 8b, respectively. Both tables show that subjects in both conditions were most likely to mention the witness evidence in their explanations than any other reason. There was no evidence of a significant difference in the percentage of subjects who mentioned the witness evidence between the two conditions (67.9% and 70.8%) ($Z = .46$, $p = .65$). When the number of witnesses decreased at Time 2 in the 15 then 10 condition, fewer subjects mentioned the witness evidence than at Time 1 ($Z = 2.02$, $p = .04$). However, in the 5 then 10 condition, subjects mentioned the witness evidence with similar frequency at Time 1 and Time 2 ($Z = .43$, $p = .67$). We also found that 15.3% of subjects at Time 2 explicitly mentioned that they thought the probability of guilt increases with a greater number of identifying witnesses. Moreover, a greater number of subjects in the 5 then 10 condition mentioned this reason than the number of subjects in the 15 then 10 condition ($Z = 3.14$, $p = .002$).

Table 8a

Reasons for Time 1 Guilt ratings in Experiment 7

Reasons	All subjects	5 witnesses	15 witnesses
Because witnesses identified the person	69.3	67.9	70.8

Mention the number of witnesses who identified the person	43.7	40.4	47.2
Because the witnesses were unanimous	56.3	53.2	59.4
Uncertainty or there was not enough evidence presented	15.3	12.8	17.9
Eyewitness evidence or memory is unreliable	10.7	11.9	9.4
The person matched the description of the perpetrator	5.6	6.4	4.7
Witnesses would have seen his face during the robbery because the perpetrator did not wear a mask	5.1	7.3	2.8

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 1 are listed. The "5 witnesses" condition presented here becomes the "5 then 10" condition in Table 8b after additional witnesses are presented. Similarly, the "15 witnesses" condition here becomes the "15 then 10" condition in Table 8b.

Table 8b

Reasons for Time 2 Guilt ratings in Experiment 7

Reasons	All subjects	5 then 10	15 then 10
Because witnesses identified the person	64.2	70.6	57.5
Mention the number of witnesses who identified the person	49.3	55.0	43.4
Because the witnesses were unanimous	38.1	40.4	35.8
Positive relationship between number of witnesses and guilt	15.3	22.9	7.5
There are still a lot of witnesses identifying the person despite the decrease	11.6	1.8	21.7

There were fewer witnesses, but they were all still unanimous	7.4	0.0	15.1
Uncertainty or there was not enough evidence presented	5.1	4.6	5.7

Note. All values are percentages. Only reasons mentioned by at least 5% of subjects at Time 2 are listed.

Discussion.

We found in Experiment 7 that if the number of witnesses identifying the suspect increased, subjects' guilt ratings increased (consistent with Experiment 1, 2, and 5), but if instead the number of witnesses decreased, there was no evidence that subject's guilt ratings decreased. These findings are consistent with Experiment 5—which was methodologically identical except for the numbers of witnesses in the conditions. In Experiment 5, we observed a small effect size (Cohen's $d = 0.05$), yet statistically significant, decrease when the number of witnesses decreased. There was no evidence of a decrease in guilt ratings in the current experiment. This discrepancy in findings is consistent with the explanation that the function of number of witnesses on guilt ratings is non-linear. That is, as the total number of witnesses identifying the suspect increases, the influence of adding or subtracting witnesses on guilt ratings decreases.

Although the difference between Experiments 5 and 7 shows that non-linear function might have contributed to the asymmetrical pattern observed in both experiments, there are likely other reasons for this pattern. The most obvious explanation for this asymmetrical pattern is that subjects are more likely to utilise evidence that confirms their belief of guilt than evidence that does not confirm that belief. This explanation is consistent with the story model of juror decision-making. Further research is needed to fully understand that mechanism. For example, are people even encoding and remembering the

details about the number of witnesses when that number decreases—or are they encoding and remembering the evidence, but dismissing it?

Experiment 7 was similar to Experiment 2 and 5 because in each of these experiments the final number of witnesses across all conditions within each experiment was identical (six witnesses in Experiments 2 and 5 and 10 witnesses in Experiment 7). In all of these experiments we found consistent differences in final guilt ratings between conditions—the effect produced a medium or high effect size in each of these experiments (Experiment 2, Cohen's $d = 0.57$ in then 2 then 6 condition; Experiment 5, Cohen's $d = .98$ in the 2 then 6 condition; Experiment 7, Cohen's $d = 0.52$ in the 5 then 10 condition). That is, by first telling subjects there were fewer witnesses than the actual number of witnesses, and asking them to determine guilt based on that smaller number of witnesses, can inflate their guilt ratings when they are subsequently asked to determine guilt for the actual number of witnesses. We achieved this manipulation in Experiment 2 by presenting witnesses as coming forward in sequential groups. We achieved this manipulation in Experiments 5 and 7 by telling subjects that there was administration error and total number of witnesses was actually higher than they were initially told. The possible practical implications of this finding are discussed in the next chapter.

Chapter 3

Summary of Findings

Across seven experiments, we conducted the first systematic investigation of how, and to what extent, people use the number of witnesses identifying the police suspect from a lineup to determine the guilt of that suspect. Subjects in our experiments were presented with lineup evidence and asked individually to review the evidence and make a determination of guilt. We found mixed evidence that the number of witnesses who identified the police suspect from a lineup influenced subjects' guilt ratings. The influence of the number of witnesses depended on the specific experimental conditions. I will, therefore, first describe the findings of each experiment individually.

In Experiment 1, subjects saw a lineup in which one member of that lineup—the police suspect—had a distinctive feature. None of the other members of the lineup had this distinctive feature. We compared conditions in which two independent unanimous witnesses identified this police suspect with conditions in which ten independent unanimous witnesses identified the police suspect. We found no evidence that the guilt ratings between these conditions were significantly different. However, when we told subjects that one day later additional independent witnesses came forward and also identified the police suspect, subjects increased their guilt ratings. We also found that explicitly warning subjects that the lineup was biased resulted in lower overall guilt ratings compared with subjects who were not warned.

In Experiment 2, we tested the possibility that these increases in guilt ratings observed in Experiment 1 could be due to either the number of distinct groups of witnesses coming forward, or the number of times subjects had to make a determination

of guilt. We found that subjects who were told six independent witnesses identified the police suspect in three separate groups of two witnesses, gave significantly higher final guilt ratings than subjects who were told six independent witnesses identified the police suspect in one group of six witnesses. We found no evidence that the number of times subjects were asked to make a determination of guilt increased those guilt ratings. Therefore, the shifts in guilt ratings observed when additional witnesses came forward were likely explained by, at least in part, the number of groups of witnesses.

In Experiment 3, we again examined the possibility of differences in guilt ratings when the witnesses were presented to subjects as one group coming forward at one time. Perhaps the lack of evidence for differences in guilt ratings from subjects who were told there were two witnesses versus those subjects who were told there were ten witnesses in Experiment 1 was either because the lineups were biased, or because the specific parameters used in Experiment 1—for example, the numbers of witnesses, the specific crime used in the scenario, or the distinctive features used. Therefore, in Experiment 3, we tested a greater range of witnesses (5 – 20) and we presented subjects with a different crime scenario and lineup material. We also tested biased and non-biased lineup administrations. We found that even between five and twenty independent witnesses coming forward in one group at one time, there was no evidence of differences in subjects' guilt ratings. Subjects in the biased administration lineups gave lower guilt ratings on average than subjects in the non-biased administration conditions.

In Experiment 4, we returned to attempting to explain the shifts in guilt ratings when additional witnesses came forward. One explanation for these shifts was that

subjects were using their first guilt ratings as an anchor for subsequent ratings. That is, subjects adjusted their first guilt ratings upwards after learning about the additional witnesses to determine their second guilt rating. In Experiment 4, we presented each subject with two cases to evaluate simultaneously. These two cases were similar, but differed in the number of witnesses who identified the police suspect (two versus six witnesses). We counterbalanced the presentation of cases to avoid order effects. We found that if we presented a case with six witnesses identifying the police suspect first, subjects gave higher guilt ratings for this case than the second presented case with two witnesses identifying the police suspect. However, when we presented a case with two witnesses identifying the police suspect first, subjects gave similar—and not significantly different—guilt ratings as the second presented case with six witnesses identifying the police suspect. This latter finding was inconsistent with Experiment 1 and Experiment 2. There were a few methodological issues with Experiment 4 that might have caused this inconsistent pattern.

In Experiment 5, we wanted to further test the anchoring and adjustment explanation for shifts in guilt ratings, but we recognised some issues with Experiment 4 that we addressed in Experiment 5. That is, first, we think subjects had difficulty remembering the details of both cases and evaluating them simultaneously. Second, we introduced a confounding variable when we justified the number of witnesses at the different cases by varying the time of day. To overcome these issues, In Experiment 5, we used one case. We either told subjects that ten or two witnesses identified the police suspect from the lineup. After subjects gave guilt ratings, we then told subjects that there had been an administrative error in the details we had just provided them.

Specifically, in both conditions, we told subjects that the actual number of witnesses who identified the police suspect was six. Subjects then gave a second guilt rating. We found that subjects who went from two to six witnesses increased their guilt ratings to a greater extent than subjects who went from ten to six witnesses decreased their guilt ratings. That is, there was evidence of an anchoring and adjustment effect in both conditions, but when the number of witnesses increased, subjects adjusted to a greater extent than if the number of witnesses decreased.

In Experiment 6 we returned to testing the possibility that there might be differences in guilt ratings when the number of independent witnesses coming forward in one group at one time differed. We thought it was possible that the number of witnesses identifying the police suspect had a non-linear effect on guilt ratings. That is, when the number of witnesses identifying the suspect is low, the addition of another corroborating witness had a greater influence on guilt ratings than when the number of witnesses identifying the suspect is high. If that were the case, the findings in Experiment 3 would suggest that this non-linear function begins to plateau before five witnesses. In Experiment 6, we re-ran Experiment 3, but instead of using groups of five, ten, fifteen, and twenty witnesses, we either told subjects that one, two, three, or four witnesses identified the police suspect. We found a non-linear function only when the lineup administration was not biased. Guilt ratings were higher when two independent witnesses identified the police suspect from the lineup compared to when there was one witness. However, beyond two witnesses in the not biased administration condition—the effect of having a greater number of witnesses who identified the police

suspect did not lead to higher guilt ratings. We found no evidence of between-subjects differences in guilt ratings in conditions in which the lineup administration was biased.

In Experiment 7, we tested another plausible explanation for the asymmetrical pattern observed in Experiment 5. Recall, in Experiment 5, subjects in the condition in which the number of witnesses went from ten to six decreased their guilt ratings to a lesser extent than subjects in the condition in which the number of witnesses went from two to six increased their guilt ratings. We thought confirmation bias might explain these findings: Subjects generally thought the suspect was guilty and therefore utilised evidence that confirmed that belief to a greater extent than evidence that disconfirmed that belief. However, the results of Experiment 6 provide another—not mutually exclusive—explanation for the asymmetrical pattern. In the 2 then 6 condition in Experiment 5, the number of witnesses is first below the point at which the function plateaus, then when the number of witnesses increases to six, it is above the point at which the function plateaus. In contrast, in the 10 then 6 condition, both numbers of witnesses are above the point in which the function plateaus. Therefore, in Experiment 7, we re-ran Experiment 5, but to attempt to isolate the effect of confirmation bias, we reduced the impact of the apparent non-linear function on the findings. To do so, we increased the number of witnesses in both conditions, so neither condition went below—or even close to—this plateau point. We ran a 15 then 10 condition and a 5 then 10 condition. We found subjects in the 15 then 10 condition did not significantly decrease their guilt ratings. We also observed a significant increase in guilt ratings in the 5 then 10 condition. However, this increase was smaller than the increase in the 2 then 6 condition observed in Experiment 5, which suggests that the non-linear function

had an impact on guilt ratings in Experiment 5. Experiment 7 also shows that the non-linear function was not the only reason for differences in guilt ratings, which means other influences, such as confirmation bias, likely influenced subjects' guilt ratings.

Between-Subjects Shifts in Guilt Ratings

I will now attempt to explain the main findings of these seven experiments, with references to relevant literature. I first turn to the finding that subjects who were told a greater number of witnesses identified the police suspect gave higher guilt ratings, but only when the number of total witnesses was one or two and the lineup or lineup administration was clearly not biased. Below, I also consider why the association between the number of witnesses and guilt ratings was constrained by these two specific conditions.

Biased lineups. Evidence in subjects' written descriptions might help us understand why the function between the number of witnesses and guilt ratings was flat for biased lineup administrations. Subjects in the biased lineup conditions were more likely to express uncertainty in their descriptions and were more likely to give a guilt rating of 50% compared to those subjects in the non-biased administration conditions. For example, in Experiment 3, 14.2% of subjects in the biased condition gave guilt rating of 50% compared to 3.3% in the non-biased administration condition. Some of these subjects also explicitly mentioned that it was a "50/50 decision" whether the person was guilty. That is, 50% probability was used to express uncertainty rather than actually representing 50% probability—consistent with de Bruin et al. (2000). It could be this uncertainty caused by the bias made a group of subjects give a guilt rating of

50% and therefore reducing the effect on guilt rating between witness conditions—we found some evidence for this in Experiment 6.

Another reason why the function between the number of witnesses and guilt ratings was flat for biased lineup administrations is perhaps subjects who saw biased lineups put less emphasis on the witness evidence, which caused smaller differences between conditions in which the number of witnesses varied. Although plausible, there are no data to support this idea in this thesis. Further research would be needed to test this idea by asking subject how much weight they put on each individual piece of evidence. However, it is also not clear to what extent subjects are actually aware of the factors that influence their guilt ratings.

Non-linear function. I now turn the finding that the number of witnesses influenced guilt ratings when the total number of witnesses was not more than two. It is not surprising that the function between number of witnesses and guilt ratings was non-linear. After all, guilt ratings are on a 0–100 scale, so at the very least, they would plateau at the ceiling value of 100% probability. However, we observed a flattening off in guilt ratings before 100. This is somewhat consistent with Gunn et al (2016), who found diagnosticity functions levelled off before reaching 100% probability for biased lineups—but these functions in Gunn et al. decreased after flattening off, which is inconsistent without our findings. Perhaps this finding is explained by the strength of evidence: Besides the witness evidence, the other evidence presented to subjects was weak circumstantial evidence. Subjects might have a threshold of certainty based on the strength of witness evidence alone. We saw some evidence for this idea in subjects' written descriptions. For example, one subject in Experiment 3 (in the non-biased 10

witness condition) gave a guilt rating of 80 and said “All of the witnesses picked the same person, which makes it pretty likely. However, witnesses can be so flawed that I still don't believe it would be a 100% chance.” In addition, in each experiment, between 5.1% and 27.0% mentioned that they were uncertain and/or they needed further evidence. It is likely this uncertainty related to witness evidence—and in the absence of any strong corroborating evidence—created this non-linear pattern, which was observed most clearly in Experiments 3 and 6.

What might have happened to the function between number of witnesses and guilt ratings if we presented stronger corroborating evidence? For example, perhaps CCTV footage was also presented at the trial, which showed the police suspect committing the crime. There are at least three possible outcomes if we introduced this extra corroborating evidence. First, the function between the number of witnesses and guilt ratings would remain the same shape but shift upwards. That is, the differences between guilt ratings on Figure 8 would remain the same but all values would be higher—and perhaps a true ceiling effect might be observed at the 100% probability value. This pattern would be consistent with the extra corroborating evidence having an additive effect on guilt rating.

The second possible result of adding extra corroborating evidence is that the function between number of witnesses and guilt ratings would change shape. That is with the corroborating evidence, like CCTV footage, subjects would require fewer unanimous witnesses to be more certain of guilt. For example, we might see a more extreme jump in guilt ratings between conditions with one witness and conditions with two witnesses. In other words, evidence of a different function shape would

demonstrate an interaction between the number of witnesses and this extra corroborating evidence.

The third possible result of adding extra corroborating evidence is that there will be no effect of the witness evidence. In the particular example of the CCTV footage, perhaps subjects would defer to this identification evidence instead of the witness evidence. That is, the presence of subjective witness evidence becomes irrelevant in the presence of a type of evidence that is perceived as stronger and more objective. If that explanation was true, then we would observe no evidence of differences in guilt ratings, regardless of the circumstances.

There is an apparent lack of literature investigating the effects of corroborating evidence and how each piece of evidence contributes towards the determination of guilt. However, there is some research demonstrating *corroboration inflation*, which is the tendency of incriminating evidence to make other evidence seem either more valid, or more incriminating (Kassin, 2012; Mosteller, 2014), which is effectively the opposite of the third possible result (null effect) described above. Corroboration inflation is often demonstrated when a seemingly reliable type of evidence, such as confession or forensic evidence, is presented with a less reliable type of evidence. Walton and Reed (2008) give a good example. Imagine a case in which a witness claims that a defendant was at the crime scene. However, the jurors don't think this witness was reliable. Then suppose forensic evidence was presented, which showed the defendant's blood was found at the crime scene. The addition of this corroborating evidence has two effects. First, it increases the probability of the jurors concluding that the defendant was at the crime scene. Second, the corroborating evidence increases the

plausibility and strength of the witness evidence—merely because they were consistent. It is this second effect that is problematic. Based on this literature, we might expect the effect of corroborating evidence, such as CCTV footage, to both increase guilt ratings and to increase the perceived strength of the witness evidence. Further research would need to be conducted to determine how corroborating evidence actually influences guilt rating functions.

Influence of unanimity. Another possible explanation for the plateau in guilt ratings after two witnesses is the influence of unanimity. That is, when two or more witnesses are present, consensus among these witnesses is more important to jurors' decisions of guilt than the actual number of witnesses. This possibility is consistent with the high percentage of subjects (between 15.5% and 56.3%) who mentioned unanimity in their guilt rating explanations. Of course, the experiments presented in this thesis cannot test this possibility because in each experiment, the witnesses unanimously identified the police suspect. One way to test the effects of unanimity would be to have conditions in which the number of unanimous witnesses identifying the police suspect is the same, and then varying the number of non-identifying witnesses. For example, we could have a condition in which there are six unanimous witnesses identifying the suspect, and another condition with six witnesses identifying the suspect, plus one non-identifying witness, and another condition with six witnesses identifying the suspect, plus two non-identifying witnesses, and so on. We might expect in such an experiment that as the number of non-identifying witnesses increases, the lower subjects' guilt ratings. However, Wells and Lindsay (1980) claimed that the criminal justice system often ignores non-identifying witnesses compared to witnesses

who do identify someone from the lineup, so perhaps we might not observe any meaningful differences between those conditions. Therefore, perhaps, in this hypothetical experiment, instead of having non-identifying witnesses, the inconsistent witness could identify a different person from the lineup.

The problem with the type of experiment suggested in the previous paragraph is that the total number of witnesses in each condition is confounded with unanimity of the witnesses. The other way to test unanimity is to keep the total number of witnesses constant across conditions, but vary the number of witnesses who identify the police suspect. For example, we could have a condition in which there are ten witnesses and they all unanimously identify the suspect, and another condition in which there are ten witnesses and nine identify the suspect, and the other witness is non-identifying (or identifies a different member of the lineup), and another condition in which there are ten witnesses and eight identify the suspect, and other two witnesses do not identify the suspect, and so on. These conditions could also be directly compared to the conditions with the same number of unanimous witnesses. For example, if unanimity was influencing subjects guilt ratings, then we would expect that subjects who are told there were eight witnesses who identified the police suspect and two witnesses did not identify the suspect would give lower guilt ratings that subjects who are told there were eight unanimous witnesses who identified the police suspect.

To some extent, the effect of unanimity would be in opposition with the effect of confirmation bias. If there is enough evidence to make most subjects believe that the police suspect is probably guilty, then we would expect subjects to seek evidence that confirms that belief. The presence of witnesses not identifying the suspect would be

non-confirmatory evidence. For example, if subjects initially believe that the suspect is guilty, they will not put as much weight on non-confirming witnesses as they do for confirming witnesses. Such a mechanism could reduce, or even eliminate, the effect of unanimity. We might, then, expect that if the other (non-witness) evidence in the scenarios is weak, then the lack of unanimity might have a greater influence than if the other evidence presented is strong because the strength of evidence should set subjects' initial beliefs about the guilt of the suspect.

Story model. The final aspect of the between-subjects effect of number of witnesses on guilt ratings is the utility of witness evidence. According to the story model, people integrate evidence to form a coherent story of what happened (Pennington & Hastie, 1981; 1986; 1988; 1993). Therefore, subjects' first ratings in all of the experiments presented in this thesis were likely based on the narratives that subjects created about the crime. It is unlikely that subjects who were told that 20 witnesses identified the police suspect would create a vastly different narrative than subjects who were told that 10 witnesses identified the police suspect. The 10 extra witnesses have little utility for subjects who are creating a narrative of what happened. Indeed, the certainty principles of *coverage* and *coherence* proposed in the story model are unlikely to differ between our witness conditions (Pennington & Hastie, 1991). Changing the number of witnesses does not make the story account for more of the presented evidence, nor does it make the story more coherent.

Within-Subjects Shifts in Guilt Ratings

Anchoring and adjustment. I will now draw on literature to help understand the finding that adding witnesses after an initial guilt rating led subjects to give higher

second guilt ratings. The first plausible reason why we might have observed an increase in guilt ratings after additional witnesses identified the suspect is because subjects might remember and update their guilt ratings from their previous decision, consistent with the literature describing the influence of prior information or decisions on future decisions (Christensen-Szalanski & Willham, 1991; Fischhoff, 1975; Mussweiler, Strack, & Pfeiffer, 2000; Tversky & Kahneman, 1974). That is, in ambiguous situations, people rely on previously presented information or previous decisions—even if this information or those decisions are irrelevant to the future decision. In the experiments presented in this thesis, we cannot classify subjects' adjustments as "insufficient" because there is no objective true guilt rating in which subjects should be adjusting towards.

Although anchoring and adjustment might explain subjects shifts in guilt ratings, we cannot be confident in this explanation without testing it. One way we could better test anchoring and adjustment is by giving subjects an arbitrary guilt rating before asking them to make a determination of guilt. For example, similar to Tversky and Kahneman's (1974) method, we could ask subjects after they had been presented with the evidence if the suspect has a higher or lower probability of being guilty than an arbitrary number, say 70% or 30%. After subjects respond, we could then ask them to determine how guilty they actually think the suspect was. We would expect subjects who were given higher anchors would give higher guilt ratings than subjects who were given lower anchors (which would demonstrate insufficient adjustment). After obtaining these first guilt ratings, we could then increase or decrease the number of witnesses (like we did in the experiments presented in this thesis). We would expect

that the anchor number presented to subjects would influence both of subjects' guilt ratings.

This anchoring explanation does not completely explain the observed within-subjects shifts in guilt ratings in our experiments because the anchoring explanation would not predict asymmetrical adjustments in subjects' guilt ratings. Recall, subjects in the conditions in which we decreased the number of witnesses did not adjust their guilt ratings to the same extent as subjects in the conditions in which we increased the number of witnesses. If the anchoring and insufficient adjustment explanation solely explained this finding, we would not expect to observe the asymmetry in changes in guilt ratings observed in our experiments.

Repetition and confirmation bias. There are other literatures that might explain these asymmetrical adjustments in guilt ratings. One possible reason is through repetition. As I outlined earlier, in conditions in which the number of witnesses identifying the suspect changed, we repeated the fact that the suspect was identified by *some* witnesses. This repetition should make the suspect more fluent to subjects, and more likely to attribute this fluency as an indicator of guilt (Hasher, Goldstein, & Toppino, 1977; Oppenheimer, 2008; Reber & Schwarz, 1999). Another possible reason for these asymmetrical adjustments could be confirmation bias. If subjects are confident that the suspect is the perpetrator based on the evidence presented to them and based on the testimony of the witnesses, they might believe the suspect is guilty. The literature on confirmation bias shows that people tend to pay attention to, remember, and use information, that is consistent with their beliefs to a greater extent than information that is inconsistent with their beliefs (Klayman & Ha, 1987; Nickerson, 1998). This

confirmation bias explanation also fits with the story model of juror decision-making which states that jurors attempt to create a narrative of the crime while evidence is presented (Mynatt, Doherty, & Tweney, 1977; Pennington & Hastie, 1992; 1993). This narrative created by jurors is influenced more by evidence presented earlier during the trial than evidence presented later in the trial. The theory states that once jurors create a narrative of the crime, they evaluate later presented evidence in terms of whether or not that evidence fits with their narrative or not. If the new evidence does not fit with their narrative, jurors evaluate and distort the evidence to fit their narrative (Carlson & Russo, 2001; Russo, 2014; Russo, Medvec & Meloy, 1996).

It is also worth mentioning the type of plausible inferences subjects would have made in conditions in which the number of witnesses increased or decreased. In both conditions, we told subjects that there had been an administrative error in the case details and then we stated the true number of witnesses. When the number of witnesses increased, it is likely that subjects assumed an omission error. That is, some witnesses were not counted in the case summary. This type of error seems plausible when witnesses are coming forward at different times because the number of witnesses originally reported was true at some point in the investigation—and perhaps the investigators failed to update that information. But in the conditions in which witnesses decreased, it is likely that subjects assumed fabrication of evidence because the number of witnesses first reported was never true. It is possible subjects could have inferred that police were trying to mislead them by presenting false information. Despite this possibility, no subjects explicitly mentioned that they made these inferences in their written descriptions. Perhaps, if we asked subjects to explain why they thought there

was an error in the case details, we could have directly measured subjects' inferences. If we did find that subjects who were told that witnesses decreased were more likely to infer that police fabricated evidence than those who were told that witnesses increased, then it makes the asymmetrical pattern of guilt ratings even more striking. That is, despite the potential greater distrust in the investigations in which the number of witnesses decreased, subjects did not reflect this distrust in their guilt ratings. Alternatively, perhaps the effect of distrust in the investigation was outweighed by the influence of repetition.

Other explanations. In addition to the anchoring and adjustment and confirmation bias explanations, there were two additional plausible explanations that we tested in Experiment 2. First was the possibility that the number of times subjects were asked to give guilt ratings influenced those ratings themselves. Second was the number of separate groups of witnesses influenced guilt ratings, regardless of the total number of witnesses. We found no evidence to support the former explanation, but we did find that presenting six witnesses as coming forward in two or three separate groups led to higher guilt ratings than if the six witnesses were presented as coming forward as one group of six at one time.

There are also at least two other plausible explanations for the within-subjects shifts that were not tested in this thesis. First, perhaps the shifts occurred due to experimenter demand. That is, subjects assumed that when the number of witnesses increased that we, as the experimenters, were expecting subjects to increase their guilt ratings. There is research demonstrating that when subjects in an experiment guess the experimental hypothesis, they respond in a way that confirms that hypothesis (Nichols

& Maner, 2008; Orne, 1962). This is particularly true when subjects have a positive attitude towards the research or the researcher. However, under some circumstances the opposite occurs: subjects respond in a way that is inconsistent with their assumed hypothesis, often because people dislike being controlled and react to the manipulation by reasserting their agency (Brehm & Brehm, 1981; Kersbergen, Whitelock, Haynes, Schroor, & Robinson, 2019; Masling, 1966). Indeed, it would seem clear to subjects that in many of our experiments we were manipulating the number of witnesses, and it would seem plausible for them to assume we were expecting an increase in guilt ratings. However, the experimenter demand could not explain the within-subjects shifts alone because demand would not predict the asymmetrical pattern in guilt ratings we found in Experiment 5 and Experiment 7. If subjects were only responding in a way that was consistent with their assumed hypothesis, we would have expected similar decreases in guilt ratings when the number of witnesses decreased. It does not make sense that subjects in one condition would be more influenced by demand than another condition because the hypothesis in both conditions should be equally obvious. Therefore, if experimenter demand did influence subjects in our experiment, it is unlikely that it was the only influence on within-subjects shifts in guilt ratings.

Another reason why experimenter demand was unlikely to be a major influence on subjects' guilt ratings in our experiments is that no subject explicitly mentioned that they were influenced by demand in the written descriptions. We might expect at least some subjects to mention demand if it had influenced them. The final aspect of demand to acknowledge is that it is not unique to an experimental environment. In real court proceedings, we might expect "prosecutor demand" or "police demand" instead of

experimenter demand. Perhaps jurors would be less likely to go along with prosecutor or police demand, compared to experimenter demand because of the potential cost associated with a wrongful conviction. However, it still remains that the presence of external pressure in this context is unavoidable.

Warning

We found that the warnings given in Experiment 1 and Experiment 2 reduced subjects guilt ratings, and the effect size of this warning was large. The effect of the warning in these experiments was much larger than previous research using judicial warnings. For example, the effects of warning jurors about the infleunce of cross-race bias on guilt judgments was much smaller than the effect size found in our experiments (O'Connor, 2013). Other research has found that subjects often struggle to comprehend judicial instructions (Baguley, McKimmie, & Masser, 2017) and many studies investigating multiple different types of judicial warnings have found that they induce juror confusion, or no evidence of any effects (Cutler, Dexter, & Penrod, 1990; Jones, Bergold, Dillon, & Penrod, 2017; Matire & Kemp, 2009; Ramirez, Zemba, & Geiselman 1996; Skalon & Beaudry, 2019). Therefore, the warning used in our experiments was much more effective than other research investigating judicial warnings.

The warning used in Experiment 1 and Experiment 2 stated that the lineup was not fair because only one person fit the description of the perpetrator. As part of this warning, we also told subjects that if we took a random group of people and gave them the police description and lineup, it is likely these people would pick the police suspect even though they themselves did not actually witness the crime. As previously discussed, our experimental procedure did not allow us to test whether this particular warning caused

subjects to be sensitive to biased lineup evidence or to be generally sceptical of the evidence and the guilt of the police suspect. To be able to test these explanations, we would need to conduct an experiment in which subjects are either warned or not warned and also either shown a biased lineup or a non-biased lineup (i.e., a 2 (warning) x 2 (bias) design). Sensitivity would be demonstrated if the warned subjects give lower guilt ratings when the lineup was biased compared to not biased. Scepticism would be demonstrated by warned subjects giving lower guilt ratings regardless of whether the lineup was biased compared to subjects who were not warned.

This differentiation between sensitivity and scepticism is important for jury trials in New Zealand because criminal proceedings in which the case against a defendant depends substantially on the correctness of witness identification evidence, the judge must warn the jury to be cautious of this evidence (Evidence Act, 2006). There is no exact wording prescribed for this warning, but the warning must do three things. First, it must warn the jury about possibility of mistaken identity. Second, it must warn the jury that mistaken witnesses might be convincing. Third, if there is more than one witness, it must warn the jury that there is a possibility that all of witnesses might be mistaken. A similar, general, warning is used in the United States, referred to as the Telfaire instruction (*United States v Telfaire*, 1972). The Telfaire instruction, unlike the New Zealand warning is designed to be read verbatim by the judge. The content of the Telfaire instruction outlines some variables that influence the reliability of the witness evidence and emphasises that jurors must find the defendant not guilty if they have reasonable doubt. Research on the Telfaire instruction is somewhat mixed, but the most realistic studies show that the instruction causes juror scepticism in eyewitness evidence, not sensitivity (Cutler, Dexter, & Penrod, 1990;

Greene, 1988; Katzev & Wishart, 1985; Laverick, 2014; Paterson, Anderson, & Kemp, 2013; Ramirez, Zemba, & Geiselman, 1996). The warning we used explained the problem with the evidence and how the problem might have influenced witnesses. This warning is different to the Telfaire instruction, which is a warning about the decision at hand.

The general warning about eyewitness evidence in New Zealand and the Telfaire instruction are different from the specific warning we gave subjects in Experiment 1 and Experiment 2. In these experiments, the warning alerted subjects to a particular problem with the witness evidence in the particular case. Such a warning, would be permitted under a different section (Section 122) of the New Zealand Evidence Act (2006), which states that if the judge is of the opinion that evidence given in the proceeding is admissible, yet unreliable, the judge can warn the jury to be cautious of whether to accept the evidence and to consider the weight to be given to that evidence. As in the general warning given about witness identification evidence, judges can choose how they warn jurors of the unreliable evidence in New Zealand. There has been some research examining the effectiveness of specific compared to general judicial warnings, which did not find conclusive evidence for the superiority of specific warnings compared to general warnings (Paterson, Anderson, & Kemp, 2013). Although, in this experiment, "specific" did not mean the warning was given because it was relevant to the specific case. Instead, the authors gave some subjects a specific example of how co-witness discussion could influence eyewitness testimony. Therefore, the results of this research are not comparable to the general and specific warnings outlined in the Evidence Act.

Perhaps it does not matter if the specific warning we used creates juror sensitivity or scepticism. If jurors are given a general warning, they must decide if the certain

circumstances of the particular case require them to either dismiss or put less weighting on identification evidence—often in accordance with their story of "what happened" (Ostrom, Werner, & Saks, 1978; Pennington & Hastie, 1992). But if jurors are given a specific warning about unreliable identification evidence, the judge has already decided that the circumstances of the particular case warrants jurors to be cautious of the evidence. Therefore, juror scepticism of the identification evidence is not necessarily an undesirable outcome when there is a specific warning about the evidence in the case. However, the specific warning relies on judges being able to accurately evaluate the reliability of a case. There is research to suggest that judges are not aware, or do not understand, many of the factors that make identification evidence unreliable (Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006). Specifically, judges' knowledge of factors that influence witness testimony was greater than potential jurors, but judges were not as knowledgeable as expert psychologists. Therefore, specific warnings rely on a different kind of sensitivity compared to general warnings: Instead of jurors being sensitive to apply caution in the appropriate circumstances; judges must be sensitive to the factors that make evidence unreliable.

If the specific warning (like the one we gave in Experiment 1 and 2) causes scepticism, what, exactly, are subjects sceptical about? Does the warning make subjects sceptical of identification evidence, regardless of the quality of that evidence, or does the warning make subjects sceptical of all of the presented evidence and the guilt of the defendant? In the literature on judicial warnings, scepticism effects are usually described with respect to the content of the warning (Cutler & Penrod, 1995; Jones, Bergold, & Penrod, 2020; Levett & Kovera, 2008; O'Donnell & Safer, 2017). That is, warnings about

identification evidence can cause scepticism about identification evidence. However, to my knowledge, there has been no direct experimental evidence that scepticism effects caused by warnings do not leak over to the other evidence presented in the case. For example, in our experiment, could a warning about the reliability of witness evidence cause a decrease in trust in the claim that the police suspect lived near the incident? To test such a possibility, a specific warning would need to be given, then subjects would need to be asked to rate the different types of evidence, rather than one holistic judgement of guilt. This research would be useful because it could distinguish between general scepticism of all evidence and scepticism about a specific type of evidence.

Finally, it is clear in Experiments 1 and 2 that in the absence of a warning, subjects gave higher guilt ratings than subjects who did receive a warning, presumably because the warned subjects put less weight on the identification evidence than subjects who were not warned. But it is also clear from Experiment 3 and Experiment 6, that warnings are not always necessary for subjects to put less weight on identification evidence. In Experiment 3 and 6, subjects in the biased conditions were told that police identified the suspect to witnesses before the witnesses made a lineup decision. The effect of simply telling people the lineup was biased in these conditions—without an explicit warning—achieved a similar effect to the explicit warning given in Experiments 1 and 2. In particular, the size of the differences between those who were warned and those who were not warned in Experiment 1 were identical to the size of the differences between those who were told the lineup was biased and those who were not told the lineup was biased in Experiment 3 (Partial $\eta^2 = .14$). What might we expect to find if a third condition was added to Experiment 3, in which the lineup was both

biased and subjects were given an explicit warning about biased lineups (compared to simply being told the lineup was biased)? Perhaps we would expect guilt ratings to be lower if subjects were warned. Lower guilt ratings in a warning condition compared to the told biased condition would indicate that some subjects in the told biased condition did not spontaneously realise the problem with the police identifying the suspect to witnesses. If we added this warning condition and observed no evidence of further decreases in guilt ratings than the told biased condition, it might indicate that the bias is obvious and all subjects are taking action without the need for an explicit warning.

Implications

This line of research on judicial warnings has clear potential implications for understanding the mechanisms that make jurors either sceptical or sensitive of evidence. I will now discuss two other implications of my research. First, our manipulation of the number of witnesses was a strength of evidence manipulation. There have been numerous studies examining strength of evidence and how that evidence influences jurors (Devine, Buddenbaum, Houp, Studebaker, & Stolle, 2009; Hepburn, 1980; Taylor & Hosch, 2004). However, much of this research used dichotomous "strong" or "weak" strength of evidence manipulations (De La Fuente, De La Fuente, & García, 2003; Eva Martin, De La Fuente, De La Fuente, & García, 2007; Kerr, Hymes, Anderson, & Weathers, 1995; Klettke, Graesser, & Powell, 2010; Leippe, Eisenstadt, Rauch, & Seib, 2004; Skolnick & Shaw, 2001; Smith & Caldwell, 1973). Our research suggests that strength of evidence manipulations should be examined over a wider range of different strengths. Although these dichotomous strength of evidence manipulations can show us if strength of evidence influences the particular dependent

variables, these manipulations cannot describe the shape and extent to which strength of evidence influences the particular dependent variables. One study using three levels of strength of evidence on determinations of guilt supported the non-linear pattern observed in our research, with a greater difference between guilt estimates in the "weak" and "moderate" conditions than the difference in the "moderate" and "strong" evidence conditions (Smith, Bull, & Holliday, 2011). Manipulating evidence strength dichotomously also does not make practical sense. For example, the strength of forensic evidence in New Zealand is evaluated using a standardised scale, according to evaluative reporting guidelines (National Institute of Forensic Science, 2017). These guidelines instruct forensic scientists to present evidence strength on a standardised numerical 11-point scale (often using likelihood ratios), accompanied with verbal descriptors, such as "moderate support for" or "very strong support against." Taken together, our research suggests that researchers should give greater consideration to measuring strength of evidence over a wider range of values.

The second implication I will discuss is how this research fits with Gunn et al (2016), who determined that the actual probability of guilt varies in a biased lineup as the number of witnesses increases. We now know that subjects' guilt ratings do not reflect the general pattern found by Gunn et al. who calculated that guilt ratings should decrease on biased lineups as the number of unanimous witnesses identifying the same person increased. We did observe that guilt ratings plateaued as the number of witnesses increased, but we found no evidence that the function decreased. Furthermore, although we did observe a difference in subjects' guilt ratings between biased and non-biased lineups, both types elicited similar functions. This finding is also

inconsistent with the probabilities calculated by Gunn et al. Therefore, the finding adds to the research showing that subjects' ratings of guilt do closely match actual guilt, determined by diagnosticity. These findings fit with the broader literature on decision-making showing that decision processes do not resemble optimal or rational strategies (Kahneman, 1991, 1994).

Limitations and Future Research

I will now discuss some of the limitations of this research. In the previous paragraphs, I mentioned two limitations: we did not investigate unanimity and there was a clear confound in Experiment 4, which made the result of experiment invalid. There are at least three other limitations. First, we did not test all of our proposed explanations for the observed within-subjects shifts in guilt ratings. One of these untested explanations was the fact that our within-subjects shifts in guilt ratings involved repeatedly identifying the police suspect. We are therefore unable to determine the exact contribution of repetition towards these shifts in guilt ratings. We could design a future experiment to test the influence of repetition by comparing two conditions. One condition would be the same as some of our current conditions in which two witnesses come forward and identify the suspect and then one day later a second group of two witnesses come forward and identify the suspect. In the second condition, four witnesses would come forward and identify the suspect and subjects would be asked to give a guilt rating. Then one day later a second police officer would present the witness evidence again and ask subjects for a second guilt rating. Differences in subjects' two guilt ratings in this second condition would demonstrate the extent to which repetition influenced guilt ratings. If subjects in the first condition

(two separate groups of two witnesses) give higher guilt ratings than the repetition condition, then that would suggest that presenting subjects as coming forward as separate groups influences guilt ratings beyond that of the influence of repeatedly identifying the suspect and giving multiple guilt ratings. If subjects in both conditions give similar guilt ratings then that might indicate the influence of having multiple groups coming forward to identify the suspect in our experiments was mostly attributable to repeatedly identifying the suspect.

The second limitation is the fact that the presentation of evidence in our set of experiments was not as complex, or realistic, as the evidence presented in real court proceedings. Indeed, we only presented subjects with witness testimony and evidence that the suspect lived close to the crime scene. In real cases, jurors are likely presented with more complex evidence. How might the complexity of evidence and the number of witnesses influence guilt ratings? In particular, how would the pattern of guilt ratings change when there is either more ambiguity or certainty that the suspect is the culprit? We could attempt to answer this question by using our general procedure and manipulating strength of evidence between conditions. For example, in some conditions, we could tell subjects that the suspect had a valid alibi at the time the crime was committed. This evidence should decrease subject's guilt ratings, but would there be equivalent decreases across differing numbers of witnesses? It is possible we might observe the pattern that we observed in the current set of experiments, but instead of guilt ratings plateauing after two witnesses, that number might be higher because subjects would require more evidence if there is an alibi.

The addition of some types of evidence might make subjects completely dismiss witness testimony—regardless of the number of witnesses. For example, if we presented subjects with a video recording of the suspect at a different location at the same time the crime was committed, we would expect subjects would not increase their guilt ratings, regardless of the number of witnesses we told them identified the suspect. In contrast, strongly incriminating evidence might also make subjects dismiss witness evidence. For example, DNA evidence from the crime scene that matches the suspect's DNA might make subjects give high guilt ratings, regardless of the number of witnesses who identify the suspect because the DNA evidence alone might cause a ceiling effect in guilt ratings. Taken together, we would expect other evidence presented during the trial to alter the extent to which subjects utilise witness testimony. Future research should be conducted to understand exactly how witness testimony from multiple witnesses interacts with other evidence presented during a court proceeding.

Another influence of not recreating a realistic courtroom experience for subjects in our experiment is the absence of other potential numerical anchors. During the course of court proceeding, numbers will be mentioned in the presence of jurors. For example, the defence or prosecutor might influence jurors by presenting numerical information, or even by telling jurors how likely they think the defendant is guilty. According to anchoring and adjustment, subjects would be swayed by this numerical information (Christensen-Szalanski & Willham, 1991; Fischhoff, 1975; Mussweiler, Strack, & Pfeiffer, 2000; Tversky & Kahneman, 1974). Therefore, judges and legal practitioners have the potential to either intentionally or unintentionally influence jurors, simply by mentioning numbers during the court proceeding. The influence of

these other potential numerical anchors might influence the effectiveness of the proposed self-generated anchors when witnesses are presented as coming forward in multiple groups.

The third and final limitation is the fact that individual juror ratings of guilt are not used in real court proceedings. Instead, jurors must deliberate to determine a verdict. What might we expect to find if subjects in our experiments had to deliberate with other subjects? Jury decision-making research suggests that the modal pre-deliberation verdict amongst jurors is often the verdict after deliberation (Kalven & Zeisel, 1966; Strasser, Kerr, & Bray, 1982; Strasser, Stella, Hanna, & Colella, 1984). Put another way, the majority of jurors most often sway the minority. Based on this research, we can infer that if a random group of 12 subjects in our experiments were asked to deliberate, that verdict would resemble the most common pre-deliberation verdict amongst the 12 subjects. The percentage of jurors with pre-deliberation guilty verdicts in our experiments was above 50 percent except for Experiment 4 (31.9% of subjects gave guilty verdicts). In Experiment 7, 84.2% of subjects gave a guilty verdict. Based on these percentages, and the research mentioned above, we would expect that if jurors did deliberate in our experiments, the majority of them would return guilty verdicts.

However, it is overly simplistic to estimate the modal pre-deliberation verdict as the jury verdict after deliberation because of the literature on the influence of minority groups on majority groups (Maass & Clark, 1984; Moscovici & Nemeth, 1974). Clearly a minority group can, and have, changed the opinions and beliefs of a majority group before. Research has uncovered some of the factors that make minority influence over a

majority more probable. Some of these factors are the size of minority and majority groups, the extent to which the minority group is perceived as consistent, reasonable, and open to compromise, and the extent to which the minority and majority group are in the same "in group" (Maass & Clark, 1984; Nemeth, 1986). For example, a group of two or three jurors could convince the other nine jurors to change their verdict from guilty to not guilty if they presented coherent alternative story of what happened. Although the example of jury deliberation has been used to demonstrate minority influence in the literature, there are few studies that have actually investigated how minority jurors influence the majority group of jurors (Clark, 1998; Salerno, 2012).

Applications

The experiments presented in this thesis were an exploratory investigation into the influence of the number of witnesses on subjects' guilt ratings; I did not intend to create a set of experiments that were ecologically valid representations of real court proceedings. That, however, does not mean there are no real-world applications of this research. One application—which is particularly concerning—is the possibility that the within-subjects manipulation in this research could be used to increase guilty verdicts. That is, prosecutors could use this research to unduly influence jurors. For example, if three witnesses identified the suspect from a lineup, the prosecutor could either present that information by telling jurors that three people identified the defendant. Or, the prosecutor could apply our research by telling jurors that the first witness viewed the lineup and identified the suspect, then tell jurors the second witness viewed the lineup and identified the witness, and then the third witness viewed the lineup and identified the suspect. If our research translates to real proceedings, we would expect this strategy

of presenting each witnesses' testimony separately to jurors will increase guilty verdicts. If that were true, this effect would add to a growing literature showing jurors are influenced by seemingly irrelevant information (Clark, 2000). For example, studies have shown that jurors are influenced by, amongst other things, the order in which evidence is presented (Lawson, 1967), the physical appearance of a defendant (Efran, 1974), and the defendant's occupation (Visher, 1987). However, the effect of presenting witnesses as coming forward in multiple groups could be more damaging than some of this previous research: The effect of adding a group of witnesses in our experiments (e.g., Cohen's $d = 0.98$ in Experiment 5) was much larger than the effect of attractiveness on guilt judgments according to a meta-analysis (Cohen's $d = 0.19$) (Mazzella & Feingold, 1994).

The above demonstrates how this research can have a negative application. Could this research be used for good? It is, perhaps, too early to know the answer to this question. More evidence and further research is required. There is potential for warnings to be used for good. But before that might happen, future research should determine what type of warnings work best when the lineup evidence is biased. Much of the judicial warning literature has failed to find evidence of juror sensitivity effects (Cutler et al., 1990; Jones, Bergold, Dillon, & Penrod, 2017; Matire & Kemp, 2009; Ramirez, Zemba, & Geiselman, 1996; Skalon & Beaudry, 2019). So, perhaps general warnings about the accuracy of identification evidence should be abandoned and replaced with specific warnings about biased lineup evidence—only when that is a relevant factor in the specific case. However, as previously discussed, this approach requires judges to be sensitive to give the warning in the correct circumstances. There

are, however, other possible solutions, for example, courts could consult expert witnesses about the quality of the identification evidence and whether a warning should be given.

Conclusions

The number of unanimous witnesses who identify a person from a lineup can influence jurors' ratings of guilt. The influence of number of witnesses had a diminishing effect on guilt ratings—after two witnesses, the addition of further witnesses had a trivial effect on guilt ratings. When independent witnesses came forward in separate groups at separate times and all identified the same person from the lineup, subjects gave higher guilt ratings than if the same number of independent witnesses came forward as one group at one time. Warning subjects that a lineup was biased led to lower guilt ratings when only one person in that lineup matched the description of the perpetrator. However, explicit warnings were not always necessary. Simply telling subjects that the person conducting the lineup identified the police suspect to witnesses prior to these witnesses making a lineup decision led to lower guilt ratings. The findings from the seven experiments raise questions about Adolf Beck's case. Although the judge deciding Beck's fate remarked about the overwhelming number of witnesses, our findings suggest that the number of witnesses might not actually influence determination of guilt as might be expected. Would the outcome of Beck's case have been the same if there were only three women who identified him in the lineup? Even with the overwhelming number of witnesses, a warning that educated the judge about the problem associated with the biased lineups (perhaps from an expert

witness) might have reduced the chances of a guilty verdict. Perhaps a warning could have prevented Beck from spending five years in prison.

References

Ackermann, A. M. (2017). *The Royal Origin of the Police Lineup: How a Queen's Funeral Changed Criminal History.* Retrieved October 5, 2020, from https://www.annmarieackermann.com/royal-origin-police-lineup

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*, 219-235. doi: 10.1177/1088868309341564

Anand, S. (2005). The Origins, Early History and Evolution of the English Criminal Trial Jury. *Atlantic Law Review*, 43, 407-432. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/alblr43&i=415

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193*, 33–35. doi: 10.2307/24943779

Asch, S. E. (1956). Studies of independence and conformity. A minority of one against a unanimous majority. *Psychological Monographs, 70,* 1-70. doi: 10.1037/h0093718

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 241-252. doi: 10.1037/0278-7393.5.3.241

Baguley, C. M., McKimmie, B. M., & Masser, B. M. (2017). Deconstructing the simplification of jury instructions: How simplifying the features of complexity

affects jurors' application of instructions. *Law and Human Behavior*, 4, 284-304. doi: 10.1037/lhb0000234

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, 27, 1069-1077. doi: 10.1177/0956797616647519

Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. London, UK: Cambridge University Press.

Benton, T. R., Ross, D. F., Bradshaw, E., Thomas, W. N., & Bradshaw, G. S. (2006). Eyewitness Memory is still not Common Sense: Comparing Jurors, Judges and Law Enforcement to Eyewitness Experts. *Applied Cognitive Psychology*, 20, 115-129. doi: 10.1002/acp.1171

Bornstein, B. H., Deffenbacher, K. A., Penrod, S. D., & McGorty, E. K. (2012). Effects of Exposure Time and Cognitive Operations on Facial Identification Accuracy: A Meta-Analysis of Two Variables Associated with Initial Memory Strength. *Psychology, Crime & Law*, 18, 473-490. doi: 10.1080/1068316X.2010.508458

Brehm, J. W., & Brehm, S. S. (1981). *Psychological reactance*. New York: Wiley.

Brigham, J. C., Meissner, C. A., & Wasserman, A. W. (1999). Applied Issues in the Construction and Expert Assessment of Photo Lineups. *Applied Cognitive Psychology*, *13*, S73-S92. doi: 10.1002/(SICI)1099-0720(199911)13:1+<S73::AID-ACP631>3.0.CO;2-4

Brooks, T. (2004). The Right to Trial by Jury. *Journal of Applied Philosophy*, 21, 197-212. doi: 10.1111/j.0264-3758.2004.00273.x

Bruer, K. C., Harvey, M. B., Adams, A. S., & Price, H. L. (2017). Judicial Discussion of Eyewitness Identification Evidence. *Canadian Journal of Behavioural Science,* 49, 209-220. doi: 10.1037/cbs0000084

Buckhout, R. (1974). Eyewitness testimony. *Scientific American*, 231, 23-31. Retrieved from https://www.jstor.org/stable/24950236

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, 7, 91-103. doi: 10.1037//1076-898X.7.2.91

Cathcart, B. (2004, October 17). The Strange Case of Adolf Beck. *The Independent*. Retrieved from https://www.independent.co.uk

Chaiken, S., & Legerwood, A. (2012). A Theory of Heuristic and Systematic Information Processing. In P. A. M. Van Lange, A. W. Kruglanski, & E.T. Higgins (Eds.), *Handbook of Theories of Social Psychology: Volume 1* (pp. 246–266). Thousand Oaks, CA: Sage.

Chen, S., Duckworth, K., & Chaiken, S. (1999). Motivated Heuristic and Systematic Processing. *Psychological Inquiry*, *10*, 44-49. doi: 10.1207/s15327965pli1001_6

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The Hindsight Bias: A Meta-Analysis. *Organizational Behavior and Human Decision Processes*, 48, 147-168. doi: 10.1016/0749-5978(91)90010-Q

Clark III, R. D. (1998). Minority influence: The role of the rate of majority defection and persuasive arguments. *European Journal of Social Psychology*, 28, 787-796. doi: 10.1002/(SICI)1099-0992(199809/10)28:5<787::AID-EJSP895>3.0.CO;2-C

Clark, J. (2000). The social psychology of jury nullification. *Law & Psychology Review,* 24, 39-58. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/lpsyr24&i=43

Clark, S. E. (2005). A Re-examination of the Effects of Biased Lineup Instructions in Eyewitness Identification. *Law and Human Behavior*, 29, 575-604. doi: 0.1007/s10979-005-7121-1

Clark, S. E. (2012). Costs and Benefits of Eyewitness Identification Reform: Psychological Science and Public Policy. *Perspectives on Psychological Science*, 7, 238-259. doi: 10.1177/1745691612439584

Clark, S. E., & Davey, S. L. (2005). The Target-to-Foils Shift in Simultaneous and Sequential Lineups. *Law and Human Behavior*, 29, 151-172. doi: 10.1007/s10979-005-2418-7

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness Identification Evidence and Innocence Risk. *Psychonomic Bulletin & Review*, *16*, 22-42. doi: 10.3758/PBR.16.1.22

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in Eyewitness Identification. *Law and Human Behavior*, *32*, 187-218. doi: 10.1007/s10979-006-9082-4

Clark, S. E., Moreland, M. B., & Rush, R. A. (2015). Lineup composition and lineup fairness. In T. Valentine & J.P Davis (Eds.), *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV* (pp. 127–157). Winchester, UK: Wiley and Sons.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed*.). Hillsdale, NJ: Erlbaum.

Collins, R. H., Walpole, S., & Edge, J. (1904). *Committee of Inquiry into the Case of Mr. Adolf Beck: Report from the Committee Together with Minutes of Evidence, Appendix and Facsimilies of Various Documents.* London, UK: Wyman & Sons, Ltd.

Colloff, M. F. (2016). *Eyewitness Identification Performance on Lineups for Distinctive Suspects* (Doctoral dissertation, University of Warwick). http://wrap.warwick.ac.uk/90153

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27*, 1227-1239. doi: 10.1177/0956797616655789

Costabile, K. A., & Klein, S. B. (2005). Finishing Strong: Recency Effects in Juror Judgments. *Basic and Applied Social Psychology*, 27, 47-58. doi: 10.1207/s15324834basp2701_5

Crosbie, M. (2020). *The Role of the Jury in a Democracy in Action*. Retrieved October 1, 2020, from https://www.districtcourts.govt.nz/about-the-courts/j/the-role-of-the-jury-in-democracy-in-action

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology and the law.* Cambridge University Press.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1989). The Eyewitness, the Expert Psychologist, and the Jury. *Law and Human Behavior*, 13, 311-332. Retrieved from https://www.jstor.org/stable/1393827

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

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification. *Law and Human Behavior*, 11, 233-258. Retrieved from http://www.jstor.org/stable/1393634

de Bruin, W. B., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty–fifty chance". *Organizational Behavior and Human Decision Processes,* 81, 115-131. doi: 10.1006/obhd.1999.2868

De La Fuente, L., De La Fuente, E. I., & García, J. (2003). Effects of pretrial juror bias, strength of evidence and deliberation process on juror decisions: New validity evidence of the Juror Bias Scale scores. *Psychology, Crime & Law*, 9, 197-209. doi: 10.1080/1068316031000116283

Desmarais, S. L., & Read, J. D. (2011). After 30 years, What do We Know About What Jurors Know? A Meta-Analytic Review of Lay Knowledge Regarding Eyewitness Factors. *Law and Human Behavior*, 35, 200-210. doi: 10.1007/s10979-010-9232-6

Devenport, J. L., Stinson, V., Cutler, B. L., & Kravitz, D. A. (2002). How Effective are the Cross-Examination and Expert Testimony Safeguards? Jurors' Perceptions of the Suggestiveness and Fairness of Biased Lineup Procedures. *Journal of Applied Psychology*, *87*, 1042-1054. doi: 10.1037/0021-9010.87.6.1042

Devine, D. J. (2012). *Jury Decision Making: The State of the Science*. New York: New York University Press.

Devine, D. J., Buddenbaum, J., Houp, S., Studebaker, N., & Stolle, D. P. (2009). Strength of evidence, extraevidentiary influence, and the liberation hypothesis: Data from the field. *Law and Human Behavior*, 33, 136-148. doi: 10.1007/s10979-008-9144-x

Devine, D. J., Olafson, K. M., Jarvis, L. L., Bott, J. P., Clayton, L. D., & Wolfe, J. M. (2004). Explaining jury verdicts: Is leniency bias for real? *Journal of Applied Social Psychology*, 34, 2069-2098. doi: 10.1111/j.1559-1816.2004.tb02691.x

Devlin, Lord P. (1982). Forward. In J. W. Shepherd, H. D. Ellis, & G. M. Davies (1982). *Identification evidence: A psychological evaluation* (pp. v-vii). Aberdeen: Aberdeen University Press.

Dicks, P. C. (2007). *The strange case of Adolf Beck: Press influence and criminal justice reform in Edwardian England* [Doctoral thesis, Marquette University]. Dissertations (1962 - 2010) Access via Proquest Digital Dissertations. https://epublications.marquette.edu/dissertations/AAI3277080

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in Police Lineups-Partial Remembering. *Journal of Police Science and Administration*, 1, 287-293.

Dreyfuss, E. (2018, August 17). A Bot Panic Hits Amazon's Mechanical Turk. *Wired*. Retrieved from https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

Efran, M. G. (1974). The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task. *Journal of Research in Personality*, 8, 45-54. doi: 10.1016/0092-6566(74)90044-0

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*. Bloomington: Indiana University, Hearing and Communication Laboratory.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and Uncertainty in Probabilistic Inference. *Psychological Review*, 92, 433. doi: 10.1037/0033-295X.92.4.433

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of Group Judgment. *Psychological Bulletin*, *84*, 158-172. doi: 10.1037/0033-2909.84.1.158

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin*, 32, 188-200. doi: 10.1177/0146167205282152

Epley, N., & Gilovich, T. (2001). Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors. *Psychological Science*, 12, 391-396. doi: 10.1111/1467-9280.00372

Eva Martin, M., De La Fuente, L., Inmaculada De La Fuente, E., & García, J. (2007). The influence of sample type, presentation format and strength of evidence on juror simulation research. *Psychology, Crime & Law*, 13, 139-153. doi: 10.1080/10683160500537431

Evans, J. S. B. (2003). In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Sciences*, *7*, 454-459. doi: 10.1016/j.tics.2003.08.012

Evans, J. S. B. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, 59, 255-278. doi: 10.1146/annurev.psych.59.103006.093629

Evidence Act 2006. Retrieved from http://www.legislation.govt.nz

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. doi: 10.3758/BF03193146

Fischhoff, B. (1975). Hindsight is Not Equal to Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty. *Journal of Experimental Psychology: Human perception and performance*, *1*, 288-299. doi: 10.1037/0096-1523.1.3.288

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect Filler Similarity in Eyewitness Lineups: A Literature Review and a Novel Methodology. *Law and Human Behavior*, 39, 62-74. doi: 10.1037/lhb0000095

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The Effect of Suspect-Filler Similarity on Eyewitness Identification Decisions: A meta-Analysis. *Psychology, Public Policy, and Law*, 19, 151-164. doi: 10.1037/a0030618

Forsyth, W. (1875). *Trial by Jury*. Jersey City, NJ: Frederick D Linn & Company

Foster, J. L. (2013). *The Consistency of Repeated Witness Testimony Leads Triers-of-Fact to Over-Rely on the Power of a Single Voice* (Doctoral Thesis, Victoria University of Wellington). http://hdl.handle.net/10063/2886

Foster, J. L., Garry, M., & Loftus, E. F. (2012). Repeated information in the courtroom. *Court Review*, 48, 44-47. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/ctrev48&i=44

Foster, R. A., Libkuman, T. M., Schooler, J. W., & Loftus, E. F. (1994). Consequentiality and Eyewitness Person Identification. *Applied Cognitive Psychology*, 8, 107-121. doi: 10.1002/acp.2350080203

Garrett, B. L. (2011). *Convicting the Innocent: Where Criminal Prosecutions Go Wrong*. Cambridge, MA: Harvard University Press.

Gelfand, A. E., & Solomon, H. (1973). A Study of Poisson's Models for Jury Verdicts in Criminal and Civil Trials. *Journal of the American Statistical Association*, 68, 271-278. doi: 10.2307/2284062

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, T. A. (1988). A Retrospective on the Criminal Trial Jury, 1200-1800. In T. A. Green and J. S. Cockburn (Eds.), *Twelve Good Men and True: The Criminal Trial Jury in England, 1200-180* (pp. 358-99). Princeton, N.J: Princeton University Press.

Greene, E. (1988). Judge's Instruction on Eyewitness Testimony: Evaluation and Revision. *Journal of Applied Social Psychology*, 18, 252-276. doi: 10.1111/j.1559-1816.1988.tb00016.x

Greene, E., & Loftus, E. F. (1984). What's new in the news? The influence of well-publicized news events on psychological research and courtroom trials. *Basic and Applied Social Psychology*, 5, 211-221. doi: 10.1207/s15324834basp0503_4

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups Versus Lineups: An Evaluation using

ROC Analysis. *Journal of Applied Research in Memory and Cognition*, 1, 221-228. doi: 10.1016/j.jarmac.2012.09.003

Gunn, L. J., Chapeau-Blondeau, F., McDonnell, M. D., Davis, B. R., Allison, A., & Abbott, D. (2016). Too good to be true: when overwhelming evidence fails to convince. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472, 1-15. doi: 10.1098/rspa.2015.0748

Hamilton, A. (1788). The Judiciary Department. In C. Rossiter (ed.), *The Federalist Papers* (491- 499). New York: Penguin

Hannaford-Agor, P. L., Hans, V. P., Mott, N. L., & Munsterman, G. T. (2002). *Are Hung Juries a Problem*. Williamsburg, VA: National Center for State Courts.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1366–1373. doi: 10.1037/a0016567

Haselton, M.G., Nettle, D., Murray, D.R. (2005). The Evolution of Cognitive Bias. In D. M. Buss, (Ed.), *The handbook of evolutionary psychology* (pp. 724–746). New York: Wiley.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16, 107-112. doi: 10.1016/S0022-5371(77)80012-1

Henderson, K. S., & Levett, L. M. (2016). Can Expert Testimony Sensitize Jurors to Variations in Confession Evidence? *Law and Human Behavior*, 40, 638-649. doi: 10.1037/lhb0000204

Hepburn, J. R. (1980). The objective reality of evidence and the utility of systematic jury selection. *Law and Human Behavior*, 4, 89-101. doi: 10.1007/BF01040485

Hoffheimer, M. H. (1989). Requiring Jury Instructions on Eyewitness Identification Evidence at Federal Criminal Trials. *Journal of Criminal Law & Criminology*, 80, 585. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/jclc80&i=597

Horowitz, I. A., & Bordens, K. S. (1988). The Effects of Outlier Presence, Plaintiff Population Size, and Aggregation of Plaintiffs on Simulated Civil Jury Decisions. *Law and Human Behavior*, *12*, 209-229. Retrieved from https://www.jstor.org/stable/1393676

Horowitz, I. A., & Bordens, K. S. (2000). The Consolidation of Plaintiffs: The Effects of Number of Plaintiffs on Jurors' Liability Decisions, Damage Awards, and Cognitive Processing of Evidence. *Journal of Applied Psychology*, 85, 909-918. doi: 10.1037/0021-9010.85.6.909

Hyman, I. E., & James Billings Jr, F. (1998). Individual Differences and the Creation of False Childhood Memories. *Memory*, 6, 1-20. doi: 10.1080/741941598

Innocence Project (n.d.). *DNA Exonerations in the United States*. Retrieved September 1, 2020, from https://innocenceproject.org/dna-exonerations-in-the-united-states

Jacowitz, K. E., & Kahneman, D. (1995). Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin*, 21, 1161-1166. doi: 10.1177/01461672952111004

Johnson, M. K., & Raye, C. L. (1981). Reality Monitoring. *Psychological Review*, 88, 67-85. doi: 10.1037/0033-295X.88.1.67

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, 114, 3-28. doi: 10.1037/0033-2909.114.1.3

Jones, A. M., Bergold, A. N., & Penrod, S. (2020). Improving juror sensitivity to specific eyewitness factors: judicial instructions fail the test. *Psychiatry, Psychology and Law*, 27, 366-385. doi: 10.1080/13218719.2020.1719379

Jones, A. M., Bergold, A. N., Dillon, M. K., & Penrod, S. D. (2017). Comparing the effectiveness of Henderson instructions and expert testimony: Which safeguard improves jurors' evaluations of eyewitness evidence? *Journal of Experimental Criminology,* 13, 29–52. doi: 10.1007/s11292-016-9279-6

Kahneman, D. (1991). Article commentary: Judgment and decision making: A personal view. *Psychological Science*, 2, 142-145. doi: 10.1111/j.1467-9280.1991.tb00121.x

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, 18-36. Retrieved from https://www.jstor.org/stable/40753012

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *Management Science*, 12, 313–327.

Kahneman, D., Schkade, D., & Sunstein, C. (1998). Shared Outrage and Erratic Awards: The Psychology of Punitive Damages. *Journal of Risk and Uncertainty*, 16, 49-86. doi: 10.1023/A:1007710408413

Kalven, H., Zeisel, H., Callahan, T., & Ennis, P. (1966). *The American Jury*. Boston: Little, Brown.

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

Kassin, S. M., Reddy, M. E., & Tulloch, W. F. (1990). Juror interpretations of ambiguous evidence. *Law and Human Behavior*, 14, 43-55. Retrieved from https://www.jstor.org/stable/1393555

Katzev, R. D., & Wishart, S. S. (1985). Impact of Judicial Commentary Concerning Eyewitness Identifications on Jury Decision Making. *Journal of Criminal Law & Criminology*, 76, 733-745. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/jclc76&i=745

Kerr, M. H., Forsyth, R. D., & Plyley, M. J. (1992). Cold Water and Hot Iron: Trial by Ordeal in England. *The Journal of Interdisciplinary History*, *22*, 573-595. doi: 10.2307/205237

Kerr, N. L., Hymes, R. W., Anderson, A. B., & Weathers, J. E. (1995). Defendant-juror similarity and mock juror judgments. *Law and Human Behavior,* 19, 545-567.

Kersbergen, I., Whitelock, V., Haynes, A., Schroor, M., & Robinson, E. (2019). Hypothesis awareness as a demand characteristic in laboratory-based eating behaviour research: An experimental study. *Appetite*, 141, 104318. doi: 10.1016/j.appet.2019.104318

Kerstholt, J. H., & Jackson, J. L. (1998). Judicial Decision Making: Order of Evidence Presentation and Availability of Background Information. *Applied Cognitive*

Psychology, 12, 445-454. doi: 10.1002/(SICI)1099-0720(199810)12:5<445::AID-ACP518>3.0.CO;2-8

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

Klerman, D. (2018). Was the Jury Ever Self Informing? In M. Mulholland & B. Pullan (Eds.), *Judicial Tribunals in England and Europe, 1200–1700* (pp. 58-80). Manchester, UK: University Press.

Klettke, B., Graesser, A. C., & Powell, M. B. (2010). Expert testimony in child sexual abuse cases: The effects of evidence, coherence and credentials on juror decision making. *Applied Cognitive Psychology,* 24, 481-494. doi: 10.1002/acp.1565

Kolodner, J. L. (1983). Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7, 243-280. doi: 10.1016/S0364-0213(83)80001-9

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, 51, 481-537. doi: 10.1146/annurev.psych.51.1.481

Kovera, M. B., McAuliff, B. D., & Hebert, K. S. (1999). Reasoning about Scientific Evidence: Effects of Juror Gender and Evidence Quality on Juror Decisions in a Hostile Work Environment Case. *Journal of Applied Psychology*, 84, 362-375. doi: 10.1037/0021-9010.84.3.362

Langbein, J. H. (2003). *The Origins of Adversary Criminal Trial*. New York: Oxford University Press.

Lawson, R. G. (1967). Order of presentation as factor in jury persuasion. *Kentucky Law Journal, 56,* 523-555. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/kentlj56&i=533

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Seib, H. M. (2004). Timing of eyewitness expert testimony, jurors' need for cognition, and case strength as determinants of trial verdicts. *Journal of Applied Psychology*, 89, 524-541.doi: 10.1037/0021-9010.89.3.524

Leverick, F. (2014). Jury directions. In J. Chalmers, F. Laverick, & A. Shaw (Eds.), *Post Corroboration Safegaurds Review Report of the Academic Expert Group* (pp. 101-117). Edinburgh, UK: The Scottish Government.

Levett, L. M., & Kovera, M. B. (2008). The effectiveness of opposing expert witnesses for educating jurors about unreliable expert evidence. *Law and Human Behavior*, 32, 363-374. doi: 10.1007/s10979-007-9113-9

Levett, L. M., Danielsen, E. M., Kovera, M. B., & Cutler, B. L. (2005). The Psychology of Jury and Juror Decision Making. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (p. 365–406). New York: The Guilford Press.

Levi, A. M., & Evans, S. (2008). Improving the Police Lineup. In S. J. Evans (Ed.), *Public Policy Issues Research Trends*, (pp. 167-210). New York, NY: Nova Science Publishers.

Lieberman, J. D. (2002). Head Over the Heart or Heart Over the Head? Cognitive Experiential Self-Theory and Extralegal Heuristics in Juror Decision

Making. *Journal of Applied Social Psychology*, *32*, 2526-2553. doi: 10.1111/j.1559-1816.2002.tb02755.x

Lieberman, J. D., & Sales, B. D. (1997). What Social Science Teaches us about the Jury Instruction Process. *Psychology, Public Policy, and Law*, 3, 589-644. doi: 10.1037/1076-8971.3.4.589

Lindsay, D. S. (1994). Memory Source Monitoring and Eyewitness Testimony. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult Eyewitness Testimony: Current Trends and Developments* (pp. 27-55). New York: Cambridge University Press.

Lindsay, R. C., & Wells, G. L. (1985). Improving Eyewitness Identifications from Lineups: Simultaneous Versus Sequential Lineup Presentation. *Journal of Applied Psychology*, *70*, 556-564. doi: 10.1037/0021-9010.70.3.556

Lindsay, R. C., Pozzulo, J. D., Craig, W., Lee, K., & Corber, S. (1997). Simultaneous Lineups, Sequential Lineups, and Showups: Eyewitness Identification Decisions of Adults and Children. *Law and Human Behavior*, 21, 391-404. doi: 10.1023/A:1024807202926

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433-442. doi: 10.3758/s13428-016-0727-z

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*, 361-366. doi: 10.1101/lm.94705

Luus, C. E., & Wells, G. L. (1991). Eyewitness Identification and the Selection of Distracters for Lineups. *Law and Human Behavior*, *15*, 43-57. Retrieved from https://www.jstor.org/stable/1394050

Maass, A., & Clark, R. D. (1984). Hidden impact of minorities: Fifteen years of minority influence research. *Psychological Bulletin*, 95, 428-450. doi: 10.1037/0033-2909.95.3.428

Malpass, R. S. (1981). Effective Size and Defendant Bias in Eyewitness Identification Lineups. *Law and Human behavior*, 5, 299-309. Retrieved from https://www.jstor.org/stable/1393669

Malpass, R. S. and Devine, P. G. (1983). Measuring the fairness of eyewitness identification lineups. In S. M. A. Lloyd-Bostock and B. R. Clifford (Eds.), *Evaluating Witness Evidence* (pp. 81-102). Chichester, UK: Wiley.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness Identification: Lineup Instructions and the Absence of the Offender. *Journal of Applied Psychology*, 66, 482-489. doi: 10.1037/0021-9010.66.4.482

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician,* 36, 158-160. doi: 10.1080/00031305.1982.10482818

Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (2017). Evaluating Lineup Fairness: Variations Across Methods and Measures. *Law and Human Behavior*, 41, 103-115. doi: 10.1037/lhb0000203

Marder, N. (2015). Jury Instructions Written for Jurors: A Perennial Challenge. In L. Solan, J. Ainsworth & R. Shuy, (Eds.), *Speaking of Language and the Law*

Conversations on the Work of Peter Tiersma, (pp. 292–296). New York, NY: Oxford University Press.

Martire, K. A., & Kemp, R. I. (2009). The Impact of Eyewitness Expert Evidence and Judicial Instruction on Juror Ability to Evaluate Eyewitness Testimony. *Law and Human Behavior*, *33*, 225-236. doi: 10.1007/s10979-008-9134-z

Martire, K. A., & Kemp, R. I. (2011). Can Experts Help Jurors to Evaluate Eyewitness Evidence? A Review of Eyewitness Expert Effects. *Legal and Criminological Psychology*, *16*, 24-36. doi: 10.1348/135532509X477225

Masling, J. (1966). Role-related behavior of the subject and psychologist and its effect upon psychological data. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 67–104). Lincoln: University of Nebraska Press.

Mazzella, R., & Feingold, A. (1994). The Effects of Physical Attractiveness, Race, Socioeconomic Status, and Gender of Defendants and Victims on Judgments of Mock Jurors: A Meta-Analysis. *Journal of Applied Social Psychology*, *24*, 1315-1338. doi: 10.1111/j.1559-1816.1994.tb01552.x

Mazzella, R., & Feingold, A. (1994). The Effects of Physical Attractiveness, Race, Socioeconomic Status, and Gender of Defendants and Victims on Judgments of Mock Jurors: A Meta-Analysis. *Journal of Applied Social Psychology*, 24, 1315-1338. doi: 10.1111/j.1559-1816.1994.tb01552.x

McQuiston, D. E., & Malpass, R. S. (2002). Validity of the Mockwitness Paradigm: Testing the Assumptions. *Law and Human Behavior*, 26, 439-453. doi: 10.1023/A:1016383305868

Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness Identification in Simultaneous and Sequential Lineups: An Investigation of Position Effects using Receiver Operating Characteristics. *Memory*, 26, 1297-1309. doi: 10.1080/09658211.2018.1464581

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous Versus Sequential Lineups. *Journal of Experimental Psychology: Applied*, 18, 361-376. doi: 10.1037/a0030609

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous Versus Sequential Lineups. *Journal of Experimental Psychology: Applied*, *18*, 361-376. doi: 10.1037/a0030609

Mitchell, K. J., & Zaragoza, M. S. (1996). Repeated exposure to suggestion and false memory: The role of contextual variability. *Journal of Memory and Language,* 35, 246-260. doi: 10.1006/jmla.1996.0014

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior*, 29, 621-637. doi: 10.1007/s10979-005-8122-9

Moscovici, S., & Nemeth, C. (1974). Social influence: II. Minority influence. In C. Nemeth (Ed.), *Social psychology: Classic and contemporary integrations*. Rand Mcnally.

Mosteller, R. P. (2015). Pernicious inferences: Double counting and perception and evaluation biases in criminal cases. *Howard Law Journal,* 58, 365-396. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/howlj58&i=385

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility. *Personality and Social Psychology Bulletin*, 26, 1142-1150. doi: 10.1177/01461672002611010

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85-95. doi: 10.1080/00335557743000053

National Institute of Forensic Science Australia New Zealand (2017). *An Introductory Guide to Evaluative Reporting.* Retrieved from https://www.anzpaa.org.au

Nemeth, C. J. (1986). Differential contributions of majority and minority influence. *Psychological Review,* 93, 23-32. doi: 10.1037/0033-295X.93.1.23

Neuschatz J.S., Wetmore S.A., Key K.N., Cash D.K., Gronlund S.D., Goodsell C.A. (2016). A Comprehensive Evaluation of Showups. In Miller M., Bornstein B. (Eds.) *Advances in Psychology and Law* (pp. 43-69). Springer, Cham. doi: 10.1007/978-3-319-29406-3_2

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology,* 135, 151-166. doi: 10.3200/GENP.135.2.151-166

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

O'Connor, M. C. (2013). Effects of Judicial Warnings About Cross-Race Eyewitness Testimony on Jurors' Judgments (Thesis, Trinity College). Trinity College Digital Repository. https://digitalrepository.trincoll.edu/theses/342

O'Donnell, C. M., & Safer, M. A. (2017). Jury instructions and mock-juror sensitivity to confession evidence in a simulated criminal case. *Psychology, Crime & Law*, 23, 946-966. doi: 10.1080/1068316X.2017.1351965

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*, 237-241. doi: 10.1016/j.tics.2008.02.014

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist, 17*, 776-783. doi: 10.1037/h0043424

Ostrom, T. M., Werner, C., & Saks, M. J. (1978). An Integration Theory Analysis of Jurors' Presumptions of Guilt or Innocence. *Journal of Personality and Social Psychology*, 36, 436-450. doi: 10.1037/0022-3514.36.4.436

Oswald, M.E., & Grosjean, S. (2004). Confirmation Bias. In R.F. Pohl (Ed.), *Cognitive illusions: A Handbook of Fallacies and Biases in Thinking, Judgment, and Memory* (pp. 79–96). Hove, England: Psychology Press.

Paterson, H. M., Anderson, D. W., & Kemp, R. I. (2013). Cautioning Jurors Regarding Co-Witness Discussion: The Impact of Judicial Warnings. *Psychology, Crime & Law*, 19, 287-304. doi: 10.1080/1068316X.2011.631539

Pennington, N., & Hastie, R. (1981). Juror Decision-Making Models: The Generalization Gap. *Psychological Bulletin*, 89, 246-287. doi: 10.1037/0033-2909.89.2.246

Pennington, N., & Hastie, R. (1981). Juror decision-making models: The generalization gap. *Psychological Bulletin*, 89, 246-287. doi: 10.1037/0033-2909.89.2.246

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51, 242-258. doi: 10.1037/0022-3514.51.2.242

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 521-533. doi: 10.1037/0278-7393.14.3.521

Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, 13, 519-558. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/cdozo13&i=543

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. Journal of Personality and Social Psychology, 62, 189–206. doi:

Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror* (pp. 192–223). New York, NY: Cambridge University Press.

Pennington, N., & Hastie, R. (1993). *The story model for juror decision-making* (pp. 192-221). Cambridge: Cambridge University Press.

Penrod, S., & Hastie, R. (1979). Models of Jury Decision Making: A Critical Review. *Psychological Bulletin*, 86, 462-492. doi: 10.1037/0033-2909.86.3.462

Petty R.E., Cacioppo J.T. (1986) The Elaboration Likelihood Model of Persuasion. In *Communication and Persuasion. Springer Series in Social Psychology* (pp. 1-24). New York, NY. Springer. doi: 10.1007/978-1-4612-4964-1_1

Pilditch, T.D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 884-889). Austin, TX: Cognitive Science Society.

Qualtrics Labs Inc. (2016). *Qualtrics survey software [internet-based software]*. Provo, UT: Qualtrics

Ramirez, G., Zemba, D., & Geiselman, R. E. (1996). Judge's Cautionary Instructions on Eyewitness Testimony. *American Journal of Forensic Psychology*, 14, 31–66. doi:

Reardon, M. (2008). *Jury Decision Making and the Story Model: How do Deliberation Style and Evidence Order Influence the Story that Juries Create?* (Doctoral dissertation, Florida International University). ProQuest Dissertations & Theses Global. https://search.proquest.com/docview/304816246?accountid=14782

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338-342. doi: 10.1006/ccog.1999.0386

Reifman, A., Gusick, S. M., & Ellsworth, P. C. (1992). Real jurors' understanding of the law in real cases. *Law and Human Behavior*, *16*, 539-554. Retrieved from https://www.jstor.org/stable/1394237

Reifman, A., Gusick, S. M., & Ellsworth, P. C. (1992). Real Jurors' Understanding of the Law in Real Cases. *Law and Human Behavior*, 16, 539-554. doi: 10.1007/BF01044622

Reynolds, J. K., & Pezdek, K. (1992). Face Recognition Memory: The Effects of Exposure Duration and Encoding Instruction. *Applied Cognitive Psychology*, 6, 279-292. doi: 10.1002/acp.2350060402

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, NJ: Lawrence Erlbaum Associates Publishers

Rumelhart, D. E., & Ortony, A. (1976). *The representation of knowledge in memory* (pp. 99-135). Center for Human Information Processing, Department of Psychology, University of California, San Diego.

Russo, J. E. (2014). The predecisional distortion of information. In E. A. Wilhelms & V. F. Reyna (Eds.), *Neuroeconomics, Judgment, and Decision Making* (pp. 91–110). New York, NY: Psychology Press.

Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes,* 66, 102-110. doi: 10.1006/obhd.1996.0041

Saks, M. J., & Kidd, R. F. (1980). Human Information Processing and Adjudication: Trial by Heuristics. *Law and Society Review*, 123-160. doi: 10.2307/3053225

Salerno, J. M. (2012). *One angry woman: Emotion expression and minority influence in a jury deliberation context* (Doctoral dissertation, University of Illinois at Chicago).

Schacter, D. L. (1999). The seven sins of memory: insights from psychology and cognitive neuroscience. *American Psychologist*, *54*, 182-203 doi: 10.1037/0003-066X.54.3.182

Schmersal, L. A. (2011). *Group Decision Making in the Jury Context: A Combined Theoretical Approach*. (Doctoral dissertation, The University of Texas). Open Access Theses & Dissertations. 2390. https://scholarworks.utep.edu/open_etd/2390

Severance, L. J., Greene, E., & Loftus, E. F. (1984). Toward Criminal Jury Instructions that Jurors can Understand. *Journal of Criminal Law & Criminology*, 75, 198-233. doi: https://heinonline.org/HOL/P?h=hein.journals/jclc75&i=210

Shapiro, P. N., & Penrod, S. (1986). Meta-Analysis of Facial Identification Studies. *Psychological Bulletin*, 100, 139-156. doi: 10.1037/0033-2909.100.2.139

Skalon, A., & Beaudry, J. L. (2019). The effectiveness of judicial instructions on eyewitness evidence in sensitizing jurors to suggestive identification procedures captured on video. *Journal of Experimental Criminology*, 16, 565-594. doi: 10.1007/s11292-019-09381-2

Skolnick, P., & Shaw, J. I. (2001). A comparison of eyewitness and physical evidence on mock-juror decision making. *Criminal Justice and Behavior,* 28, 614-630. doi: 10.1177/009385480102800504

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair Lineups are Better than Biased Lineups and Showups, but not Because they Increase Underlying Discriminability. *Law and Human Behavior*, 41, 127-145. doi: 10.1037/lhb0000219

Smith, L. L., Bull, R., & Holliday, R. (2011). Understanding juror perceptions of forensic evidence: Investigating the impact of case context on perceptions of forensic

evidence strength. *Journal of Forensic Sciences*, 56, 409-414. doi: 10.1111/j.1556-4029.2010.01671.x

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. New York: Routledge

Stasser, G and Kerr, Norbert L. and Bray, R (1982) The social psychology of jury deliberations: Structure, process, and product. In N. L. Kerr, and R. Bray, (Eds.), *The psychology of the courtroom*. Academic Press, New York.

Stasser, G., Stella, N., Hanna, C., & Colella, A. (1984). The majority effect in jury deliberations: Number of supporters versus number of supporting arguments. *Law & Psychology Review,* 8, 115-128. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/lpsyr8&i=117

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-Two Tests of the Sequential Lineup Superiority Effect: A Meta-Analysis and Policy Discussion. *Psychology, Public Policy, and Law*, 17, 99-139. doi: 10.1037/a0021650

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness Accuracy Rates in Sequential and Simultaneous Lineup Presentations: A Meta-Analytic Comparison. *Law and Human Behavior*, 25, 459-473. doi: 10.1023/A:1012888715007

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness Accuracy Rates in Sequential and Simultaneous Lineup Presentations: A Meta-Analytic Comparison. *Law and Human Behavior*, *25*, 459-473. doi: 10.1023/A:1012888715007

Sue, S., Smith, R. E., & Caldwell, C. (1973). Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. *Journal of Applied Social Psychology*, 3, 345-353. doi: 10.1111/j.1559-1816.1973.tb02401.x

Sweeney, L. T., & Haney, C. (1992). The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, 10, 179-195. doi: 10.1002/bsl.2370100204

Tapp, J. L., & Levine. (1973). The Psychology of Criminal Identification: The Gap from Wade to Kirby. *University of Pennsylvania Law Review,* 121, 1079-1084.

Taylor, T. S., & Hosch, H. M. (2004). An examination of jury verdicts for evidence of a similarity-leniency effect, an out-group punitiveness effect or a black sheep effect. *Law and Human Behavior*, 28, 587-598. doi: 10.1023/B:LAHU.0000046436.36228.71

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, 22, 217-237. doi: 10.1023/A:1025746220886

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 1124-1131. doi: 10.1126/science.185.4157.1124

Unkelbach, C., Fiedler, K., & Freytag, P. (2007). Information repetition in evaluative judgments: Easy to monitor, hard to control. *Organizational Behavior and Human Decision Processes*, 103, 37-52. doi: 10.1016/j.obhdp.2006.12.002

Visher, C. A. (1987). Juror decision-making. *Law and Human Behavior, 11*, 1-17. doi: 10.1007/BF01044835

von Moschzisker, R. (1921). Historic Origin of Trial by Jury. *University of Pennsylvania Law Review and American Law Register*, 70, 73-86. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/pnlr70&i=107

Walton, D., & Reed, C. (2008). Evaluating corroborative evidence. *Argumentation*, 22, 531-553. doi: 10.1007/s10503-008-9104-0

Wang, W. C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. *Journal of Cognitive Neuroscience*, 28, 739-746. doi: 10.1162/jocn_a_00923

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology,* 92, 821-833. doi: 10.1037/0022-3514.92.5.821

Wells, G. L. (1984). The Psychology of Lineup Identifications. *Journal of Applied Social Psychology*, 14, 89-103. doi: 10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 88, 776-784. doi: 10.1037/0033-2909.88.3.776

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence. *Law and Human Behavior*, 44, 3-36. doi: 10.1037/lhb0000359

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for Empirically Assessing the Fairness of a Lineup. *Law and Human Behavior*, 3, 285-293. doi: 10.1007/BF01039807

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The Selection of Distractors for Eyewitness Lineups. *Journal of Applied Psychology*, *78*, 835-844. doi: 10.1037/0021-9010.78.5.835

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness Identification Procedures: Recommendations for Lineups and Photospreads. *Law and Human Behavior*, 22, 603-647. Retrieved from https://www.jstor.org/stable/1394446

West, E., & Meterko, V. (2015). Innocence project: DNA exonerations, 1989-2014: review of data and findings from the first 25 years. *Albany Law Review,* 79, 717–795. Retrieved from http://www.albanylawreview.org/Pages/home.aspx

White, R. H. (1961). Origin and Development of Trial by Jury. *Tennessee Law Review*, 29, 8-18. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/tenn29&i=24

Winter, R. J., & Greene, E. (2007). Juror Decision-Making. *Handbook of Applied Cognition*, 2, 739-762. doi: 10.1002/9780470713181

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. Empirical Discriminability: The Application of ROC Methods to Eyewitness Identification. *Cognitive Research: Principles and Implications*, 3:9, 1-29. doi: 10.1186/s41235-018-0093-8

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of Lineup Memory. *Cognitive Psychology*, *105*, 81-114. doi: 10.1016/j.cogpsych.2018.06.001

Wogalter, M. S., Malpass, R. S., & McQuiston, D. E. (2004). A National Survey of US Police on Preparation and Conduct of Identification Lineups. *Psychology, Crime and Law*, 10, 69-82. doi: 10.1080/10683160410001641873

Woods, L. D. (2018). The Miseducation of the American Juror. *J. Race Gender & Poverty*, *10*, 19-38. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/jrgenpo10&i=25

Wright, D. S., Wade, K. A., & Watson, D. G. (2013). Delay and déjà vu: Timing and repetition increase the power of false evidence. *Psychonomic Bulletin & Review*, 20, 812-818. doi: 10.3758/s13423-013-0398-z

Wykes, T. G. (2014). *Juror perceptions of eyewitness identification evidence* (Masters thesis, Wilfrid Laurier University). Theses and Dissertations (Comprehensive). https://scholars.wlu.ca/etd/1679

Zaragoza, M. S., & Mitchell, K. J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science,* 7, 294-300. doi: 10.1111/j.1467-9280.1996.tb00377.x

Appendix A

Instructions, perpetrator descriptions, and lineups presented to subjects in Experiment 1.

Two weeks ago, Robert Brown was in a downtown parking lot, calling his friend from a cell phone. During this phone conversation, another person approached Robert and snatched his phone out of his hands. The perpetrator then fled the parking lot on foot. [***Two or Ten***] **people** witnessed the perpetrator steal the phone. The police created this description of the perpetrator based on what the [*two or ten*] witnesses told them:

Young male (early 20's)

5' 10" tall

Short hair

[*Large scar on his face or Black eye*]

One week after the crime, police identified a man who lived close to the parking lot as a suspect, and considered him their prime suspect. Police then created a photo lineup using a picture of the suspect and five other people. Each witness was asked if the person who took Robert's phone was in the lineup. Witnesses were not told about the choices the other witness made.

Here is the lineup the witnesses saw: [shown one of the lineups below]

Figure A1. Lineup 1: Left Scar

Figure A2. Lineup 2: Right Scar

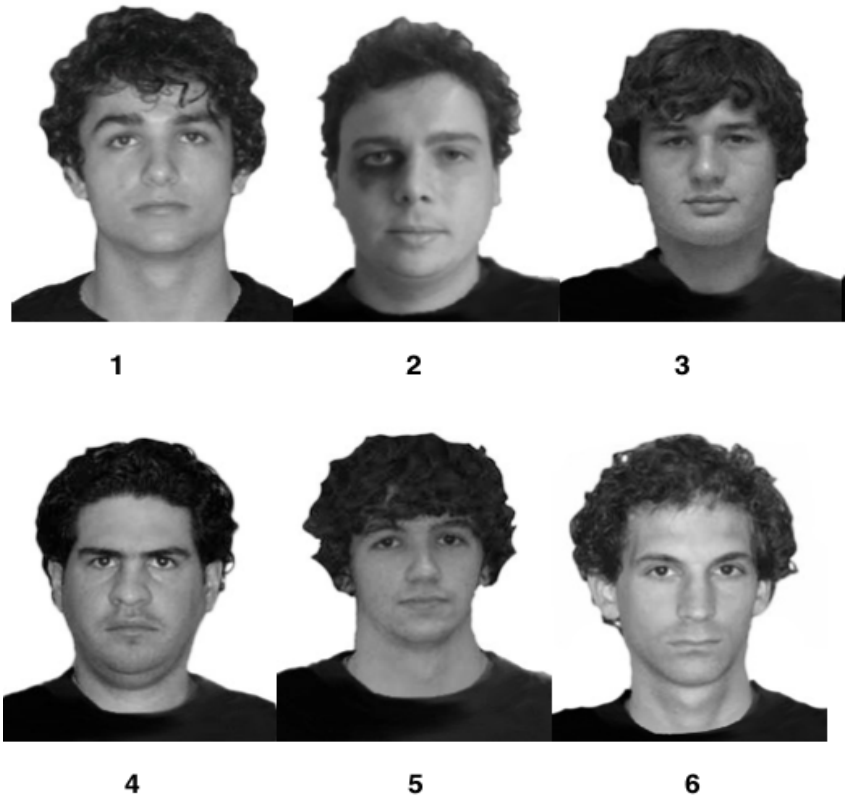

Figure A3. Lineup 3: Black Eye

[Then subjects in each condition was told:]

The [*two or ten*] witnesses picked Person x [*the person in the lineup that had the distinctive feature*] as the person who stole Robert's phone

Person x was also the police suspect.

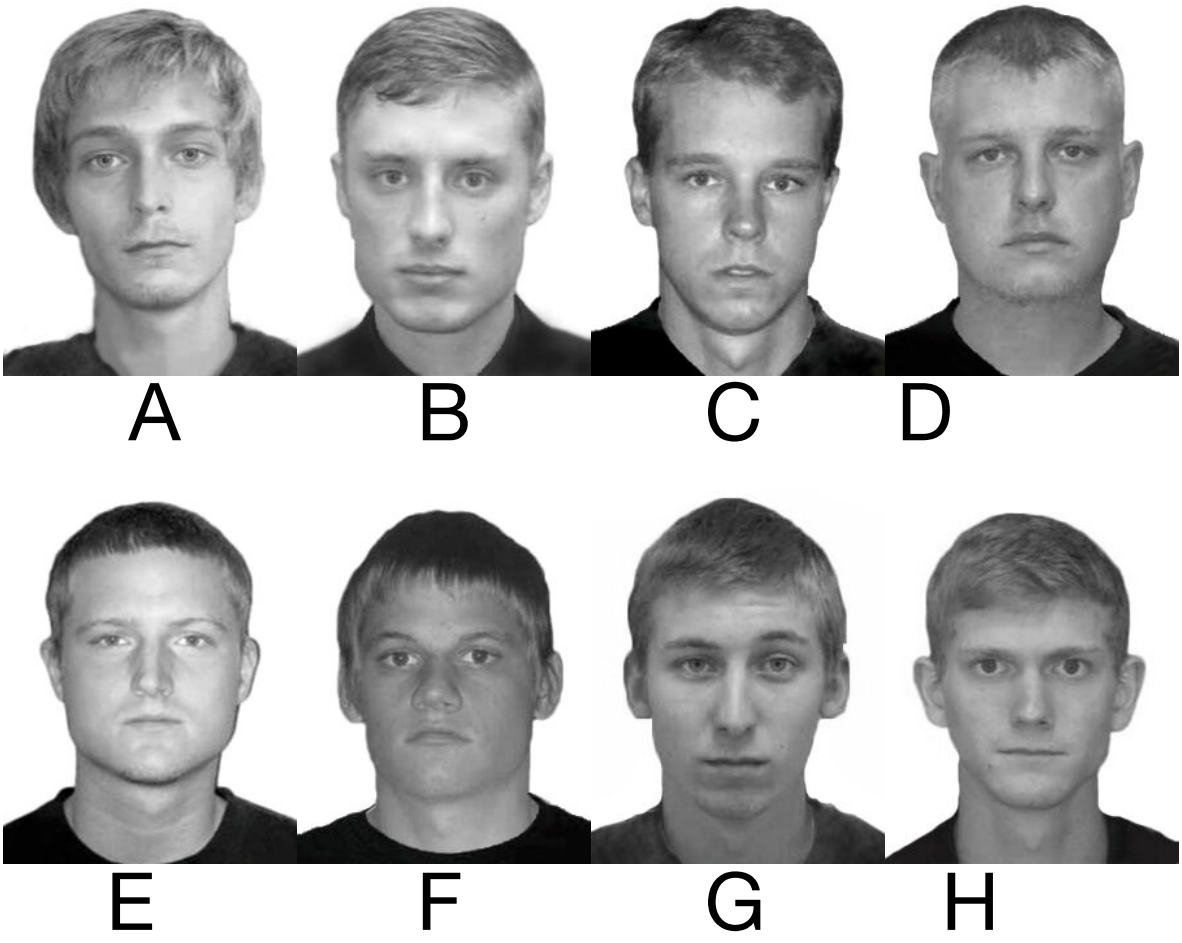

Figure A4. Lineup 4 used in Experiments 3, 4, 5, 6, and 7

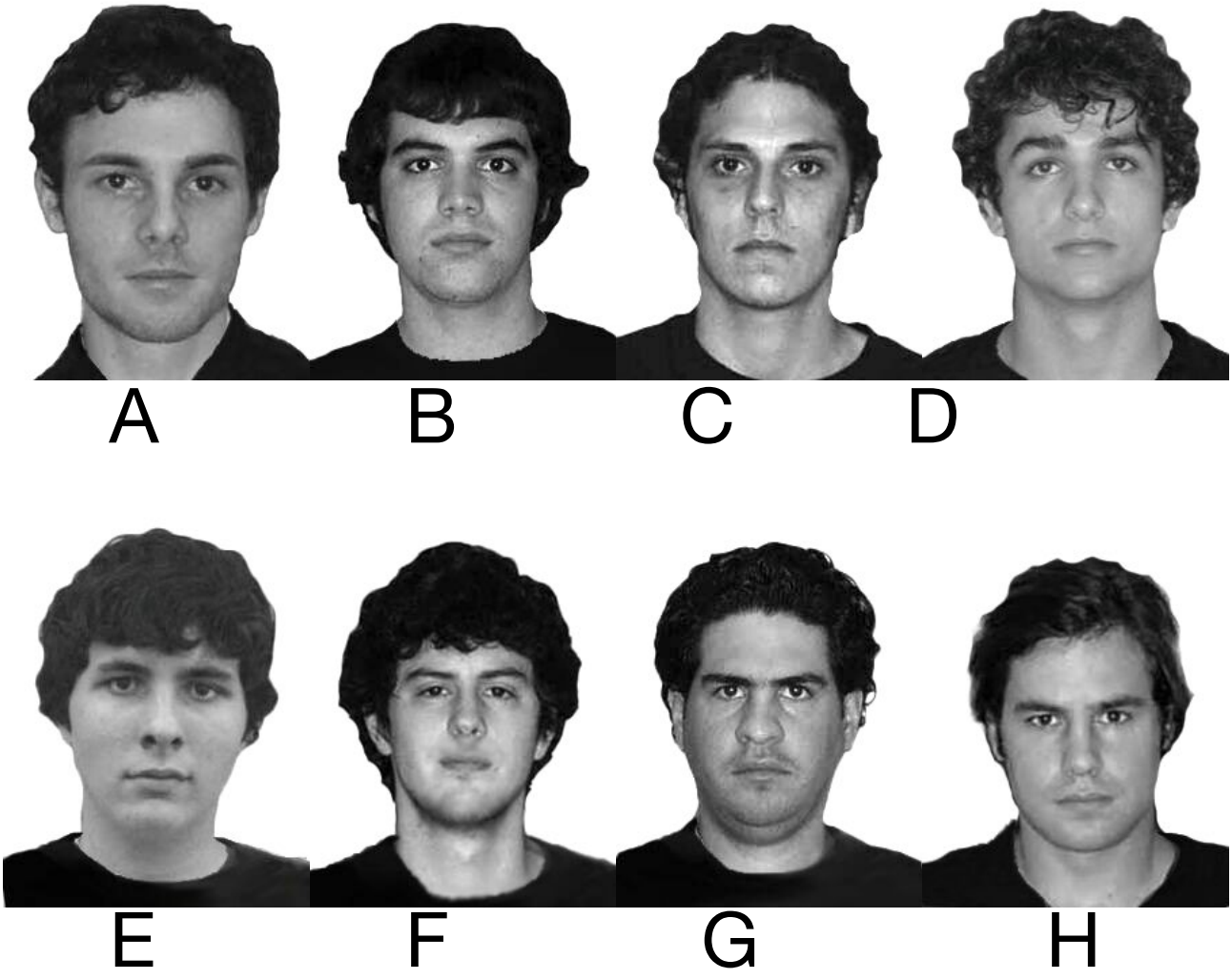

Figure A5. Lineup 4 used in Experiments 3, 4, 5, 6, and 7

Appendix B

Table B1

Compliance Questions Presented to Subjects and the Percentage of Subjects who Failed each Question

Question	E1	E2	E3a	E3b	E4	E5	E6a	E6b	E7
Did you maximize the size of your web browser so that it covers your entire screen?	8.0	5.5	2.1	5.4	3.9	5.9	9.7	5.9	6.0
Did you complete the experiment on a mobile phone (or a similar device with a small screen)?	6.1	6.4	4.2	3.3	10.8	11.2	18.8	8.3	10.2
Did you complete the experiment in a single session, without stopping?	1.0	2.0	1.7	0.8	1.5	0.5	3.4	3.4	2.3
Did you pause or leave the experiment to engage in other tasks, even if they were other computer tasks?	1.7	3.1	1.3	2.5	6.9	9.3	27.5	14.6	10.2
Did you complete the experiment without anyone helping you?	6.9	3.5	3.3	4.5	2.9	6.3	11.6	10.2	7.9
Did you complete the experiment in an environment that is free of noise and distraction?	8.5	6.8	3.3	4.1	2.5	4.9	13.5	6.8	6.0
Did you speak with anyone at any time during the experiment?	2.6	4.0	0.8	2.5	6.9	10.7	26.1	11.7	7.4
Subjects who failed at least one of the above	**27.0**	**21.0**	**12.1**	**17.4**	**20.6**	**30.2**	**49.3**	**32.7**	**27.0**

Note. The bottom row is not equivalent to the sum of each column because some subjects failed multiple questions.

Table B2

The Percentage of Subjects who Incorrectly Answered the Attention Questions from each Experiment

Question	E1	E2	E3a	E3b	E4	E5	E6a	E6b	E7
According to what you read, how many witnesses told the police they saw the crime?	29.9	25.0	15.4	12.5	44.1	26.8	40.1	22.0	19.5
According to what you read, how many of these witnesses picked Person 4 from the lineup?	29.9	25.2	14.6	10.4	62.3	25.9	34.3	21.5	23.3
Answered at least one question incorrectly	36.2	30.1	19.2	14.2	65.2	33.2	48.3	28.3	27.0

Note. The bottom row is not equivalent to the sum of each column because some subjects failed multiple questions. In Experiments 1, 2, 4, 5, and 7 the number of witnesses changed and subjects were asked these questions multiple times. In these experiments, we coded incorrect responses as at least one incorrect response.

Appendix C

Supplementary Analyses for Experiment 1

Primary Analysis

We ran a 2 (Warning: warning; no warning) x 2 (Witness Condition: 2 then 12; 10 then 12) x 2 (Guilt Rating: Time 1; Time 2) x 3 (Distinctive Feature: left scar, right scar, black eye) mixed ANOVA, with guilt rating as a repeated measure and warning, witness condition, and distinctive feature as between-subjects factors.

The Distinctive Feature x Witness Condition x Warning x Guilt Rating four-way interaction was not significant: $F(2, 566) = .17$ $p = .84$, Partial $\eta^2 = .001$.

The Guilt Rating x Witness Condition x Warning three-way interaction was not significant: $F(1, 566) = .93$, $p = .34$, Partial $\eta^2 = .002$.

The Guilt Rating x Distinctive Feature x Warning three-way interaction was not significant: $F(2, 566) = .02$, $p = .98$, Partial $\eta^2 < .001$.

The Guilt Rating x Distinctive Feature x Witness Condition three-way interaction was not significant: $F(2, 566) = 6.48$, $p = .002$, Partial $\eta^2 = .02$.

The Distinctive Feature x Witness Condition x Warning three-way interaction was not significant: $F(2, 566) = 2.33$, $p = .098$, Partial $\eta^2 = .01$.

The Guilt Rating x Warning two-way interaction was significant: $F(1, 566) = 4.97$, $p = .03$, Partial $\eta^2 = .01$.

The Guilt Rating x Witness Condition two-way interaction was significant: $F(1, 566) = 9.22$, $p = .003$, Partial $\eta^2 = .02$.

The Guilt Rating x Distinctive Feature two-way interaction was not significant: $F(2, 566) = 1.88$, $p = .15$, Partial $\eta^2 = .01$.

The Witness Condition x Warning two-way interaction was not significant: $F(1, 566) = .01$, $p = .93$, Partial $\eta^2 = .000$.

The Distinctive Feature x Warning two-way interaction was not significant: $F(2, 566) = .42$, $p = .66$, Partial $\eta^2 = .001$.

The Distinctive Feature x Witness Condition two-way interaction was not significant: $F(2, 566) = .21$, $p = .81$, Partial $\eta^2 = .001$.

The warning main effect was significant: $F(1, 566) = 91.22$, $p < .001$, Partial $\eta^2 = .14$.

The witness condition main effect was not significant: $F(1, 566) = .48$, $p = .49$, Partial $\eta^2 = .001$.

The distinctive feature main effect was not significant: $F(2, 566) = 1.24$, $p = .29$, Partial $\eta^2 = .004$.

The guilt rating main effect was significant: $F(1, 566) = 315.60$, $p < .001$ Partial $\eta^2 = .36$.

Compliance Check Exclusions

We ran a 2 (Warning: warning; no warning) x 2 (Witness Condition: 2 then 12; 10 then 12) x 2 (Guilt Rating: time 1; time 2) x 3 (Distinctive Feature: left scar, right scar, black eye) mixed ANOVA, with guilt rating as a repeated measure and warning, witness condition, and distinctive feature as between-subjects factors. We excluded the 21.0% of subjects who failed at least one compliance check question.

The Distinctive Feature x Witness Condition x Warning x Guilt Rating interaction was not significant: $F(2, 410) = .38$ $p = .69$, Partial $\eta^2 = .00$.

The Guilt Rating x Witness Condition x Warning interaction was not significant: $F(1, 410) = 2.62$, $p = .11$, Partial $\eta^2 = .01$.

The Guilt Rating x Distinctive Feature x Warning interaction was not significant: $F(2, 410) = .20$, $p = .82$, Partial $\eta^2 = .001$.

The Guilt Rating x Distinctive Feature x Witness Condition interaction was not significant: $F(2, 410) = 4.19$, $p = .02$, Partial $\eta^2 = .02$.

The Distinctive Feature x Witness Condition x Warning interaction was not significant: $F(2, 410) = 1.45$, $p = .24$, Partial $\eta^2 = .01$.

The Guilt Rating x Warning interaction was significant: $F(1, 410) = 6.83$, $p = .01$, Partial $\eta^2 = .02$.

The Guilt Rating x Witness Condition interaction was not significant: $F(1, 410) = 3.27$, $p = .07$, Partial $\eta^2 = .01$.

The Guilt Rating x Distinctive Feature interaction was not significant: $F(2, 410) = 1.79$, $p = .17$, Partial $\eta^2 = .01$.

The Witness Condition x Warning interaction was not significant: $F(1, 410) = .03$, $p = .86$, Partial $\eta^2 = .000$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 410) = .82$, $p = .44$, Partial $\eta^2 = .004$.

The Distinctive Feature x Witness Condition interaction was not significant: $F(2, 410) = .03$, $p = .97$, Partial $\eta^2 = .000$.

The warning main effect was significant: $F(1, 410) = 80.52$, $p < .001$, Partial $\eta^2 = .16$.

The witness condition main effect was not significant: $F(1, 410) = .77$, $p = .38$, Partial $\eta^2 = .002$.

The distinctive feature main effect was not significant: $F(2, 410) = 1.53$, $p = .22$, Partial $\eta^2 = .007$.

The guilt rating main effect was significant: $F(1, 410) = 249.62$, $p < .001$ Partial $\eta^2 = .38$.

Attention Check Exclusions

We ran a 2 (Warning: warning; no warning) x 2 (Witness Condition: 2 then 12; 10 then 12) x 2 (Guilt Rating: time 1; time 2) x 3 (Distinctive Feature: left scar, right scar, black eye) mixed ANOVA, with guilt rating as a repeated measure and warning, witness condition, and distinctive feature as between-subjects factors. We excluded the 30.1% of subjects who failed at least one attention check question.

The Distinctive Feature x Witness Condition x Warning x Guilt Rating interaction was not significant: $F(2, 357) = .17$ $p = .84$, Partial $\eta^2 = .001$.

The Guilt Rating x Witness Condition x Warning interaction was not significant: $F(1, 357) = .41$, $p = .52$, Partial $\eta^2 = .001$.

The Guilt Rating x Distinctive Feature x Warning interaction was not significant: $F(2, 357) = .01$, $p = .99$, Partial $\eta^2 = .000$.

The Guilt Rating x Distinctive Feature x Witness Condition interaction was significant: $F(2, 357) = 3.31$, $p = .04$, Partial $\eta^2 = .02$.

The Distinctive Feature x Witness Condition x Warning interaction was not significant: $F(2, 357) = .38$, $p = .68$, Partial $\eta^2 = .002$.

The Guilt Rating x Warning interaction was significant: $F(1, 357) = 6.71$, $p = .01$, Partial $\eta^2 = .02$.

The Guilt Rating x Witness Condition interaction was significant: $F(1, 357) = 15.93$, $p < .001$, Partial $\eta^2 = .04$.

The Guilt Rating x Distinctive Feature interaction was not significant: $F(2, 357) = .19$, $p = .83$, Partial $\eta^2 = .001$.

The Witness Condition x Warning interaction was not significant: $F(1, 357) = 1.65$, $p = .20$, Partial $\eta^2 = .01$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 357) = .82$, $p = .44$, Partial $\eta^2 = .01$.

The Distinctive Feature x Witness Condition interaction was not significant: $F(2, 357) = 1.48$, $p = .23$, Partial $\eta^2 = .001$.

The warning main effect was significant: $F(1, 357) = 60.68$, $p < .001$, Partial $\eta^2 = .15$.

The witness condition main effect was not significant: $F(1, 357) = .40$, $p = .53$, Partial $\eta^2 = .001$.

The distinctive feature main effect was not significant: $F(2, 357) = .49$, $p = .61$, Partial $\eta^2 = .003$.

The guilt rating main effect was significant: $F(1, 357) = 233.88$, $p < .001$ Partial $\eta^2 = .40$.

Lineup Fairness Ratings

We ran a 2 (Warning: warning; no warning) x 2 (Witness Condition: 2 then 12; 10 then 12) x 3 (Distinctive Feature: scar 1; scar 2; black eye) ANOVA, with lineup fairness ratings as the dependent variable. The significant main effect of warning was presented in Chapter 2 ($F(1, 566) = 121.15$, $p < .001$, Partial $\eta^2 = .18$). The remaining, non-significant, main effects and interactions are presented below.

The distinctive feature main effect was not significant: $F(2, 566) = 1.60$, $p = .21$, Partial $\eta^2 = .01$.

The witness condition main effect was not significant: $F(1, 566) = .25$, $p = .61$, Partial $\eta^2 < .001$.

The Distinctive Feature x Witness Condition interaction was not significant: $F(2, 566) = .38$, $p = .69$, Partial $\eta^2 = .001$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 566) = .77$, $p = .47$, Partial $\eta^2 = .003$.

The Witness Condition x Warning interaction was not significant: $F(1, 566) = .40$, $p = .53$, Partial $\eta^2 = .001$.

The Distinctive Feature x Witness Condition x Warning interaction was not significant: $F(2, 566) = .71$, $p = .49$, Partial $\eta^2 = .003$.

Supplementary Analyses for Experiment 2

Primary Analysis

We ran a 3 (Distinctive Feature: left scar; right scar; black eye) x 3 (Number of Witness groups: 1; 2; 3) x 2 (Warning: warned; not warned) ANOVA with final guilt rating as the dependent variable. The three main effects from this model are displayed in the main text. Here, we report the non-significant interactions.

The Distinctive Feature x Number of Witness Groups x Warning interaction was not significant: $F(4, 530) = 1.11$, $p = .35$, Partial $\eta^2 = .01$.

The Number of Witness Groups x Warning interaction was not significant: $F(2, 530) = .44$, $p = .64$, Partial $\eta^2 = .002$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 530) = .17$, $p = .85$, Partial $\eta^2 = .001$.

The Distinctive Feature x Number of Witness Groups interaction was not significant: $F(4, 530) = .51$, $p = .73$, Partial $\eta^2 = .004$.

Compliance Check Exclusions

We ran a 3 (Distinctive Feature: left scar; right scar; black eye) x 3 (Number of Witness Groups: 1; 2; 3) x 2 (Warning: warned; not warned) ANOVA, excluding the 21.0% who failed the compliance checks.

The Distinctive Feature x Number of Witness Groups x Warning interaction was not significant: $F(4, 415) = 1.10$, $p = .36$, Partial $\eta^2 = .01$.

The Number of Witness groups x Warning interaction was not significant: $F(2, 415) = .46$, $p = .63$, Partial $\eta^2 = .002$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 415) = .05$, $p = .95$, Partial $\eta^2 = .000$.

The Distinctive Feature x Number of Witness Groups interaction was not significant: $F(4, 415) = .92$, $p = .45$, Partial $\eta^2 = .01$.

The warning main effect was significant: $F(1, 415) = 48.02$, $p < .001$, Partial $\eta^2 = .10$. Subjects who were not warned the lineup was biased gave higher final guilt ratings ($M = 84.27$, $SD = 18.24$) than subjects who were warned the lineup was biased ($M = 69.13$, $SD = 26.13$).

The number of witness groups main effect was significant: $F(2, 415) = 7.15$, $p = .001$, Partial $\eta^2 = .03$. Tukey post-hoc tests showed subjects in the three groups condition ($M = 76.54$, $SD = 24.80$) gave significantly higher final guilt ratings than subjects in the one group condition ($M = 68.52$, $SD = 25.70$) ($p = .02$). Subjects in the two groups condition ($M = 79.62$, $SD = 21.43$) gave significantly higher final guilt ratings than subjects in the one group condition ($p = .001$). No other differences were statistically significant.

The distinctive feature main effect was not significant: $F(2, 415) = .31$, $p = .74$, Partial $\eta^2 = .001$.

We ran a one-way ANOVA, comparing subjects in the three witness groups condition and two witness groups condition, to determine if final guilt ratings differed by whether subjects gave one guilt rating after all of the evidence had been presented, or whether subjects gave guilt ratings after each groups of witnesses came forward. We excluded the 21.0% who failed the compliance checks. The ANOVA was not significant, $F(1, 349) = 0.29$, $p = .59$.

Attention Check Exclusions

We ran a 3 (Distinctive Feature: left scar; right scar; black eye) x 3 (Number of Witness Groups: 1; 2; 3) x 2 (Warning: warned; not warned) ANOVA, excluding the 30.1% who could not recall at least one of the questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Distinctive Feature x Number of Witness Groups x Warning interaction was not significant, $F(4, 365) = 1.94$, $p = .10$, Partial $\eta^2 = .02$.

The Number of Witness Groups x Warning interaction was not significant, $F(2, 365) = .04$, $p = .96$, Partial $\eta^2 = .000$.

The Distinctive Feature x Warning interaction was not significant, $F(2, 365) = .46$, $p = .64$, Partial $\eta^2 = .002$.

The Distinctive Feature x Number of Witness Groups t interaction was not significant, $F(4, 365) = .33$, $p = .86$, Partial $\eta^2 = .004$.

The warning main effect was significant: $F(1, 365) = 50.44$, $p < .001$, Partial $\eta^2 = .12$. Subjects who were not warned the lineup was biased gave higher final guilt ratings ($M = 86.13$, $SD = 16.51$) than subjects who were warned the lineup was biased ($M = 69.05$, $SD = 25.91$).

The number of witness groups main effect was significant: $F(2, 365) = 5.02$, $p = .007$, Partial $\eta^2 = .03$. Tukey post-hoc tests showed subjects in the three witness groups condition ($M = 78.51$, $SD = 23.96$) gave significantly higher final guilt ratings than subjects in the one witness group condition ($M = 69.96$, $SD = 25.31$) ($p = .01$). Subjects in the two witness groups condition ($M = 79.71$, $SD = 21.36$) gave significantly higher final

guilt ratings than subjects in the one witness group condition (p = .004). No other differences were statistically significant.

The Distinctive Feature main effect was not significant: $F(2, 365) = .83$, $p = .44$, Partial $\eta^2 = .01$.

We ran a one-way ANOVA, comparing subjects in the three witness groups condition and two witness groups condition, to determine if final guilt ratings differed by whether subjects gave one guilt rating after all of the evidence had been presented, or whether subjects gave guilt ratings after each groups of witnesses came forward. We excluded the 30.1% who could not recall at least one of the questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup. The ANOVA was not significant, $F(1, 307) = 1.81$, $p = .18$.

Lineup Fairness Ratings

We ran a 3 (Distinctive Feature: scar 1; scar 2; black eye) x 3 (Number of Witness Groups: 1; 2; 3) x 2 (Warning: warning; no warning) ANOVA, with lineup fairness ratings as the dependent variable. The significant main effect of warning was presented in Chapter 2 ($F(1, 530) = 131.91$, $p < .001$, Partial $\eta^2 = .20$). The remaining, non-significant, main effects and interactions are presented below.

The distinctive feature main effect was not significant: $F(2, 530) = .19$, $p = .83$, Partial $\eta^2 = .001$.

The number of witness groups main effect was significant: $F(2, 530) = 3.08$, $p = .05$, Partial $\eta^2 = .01$.

The Distinctive Feature x Number of Witness Groups interaction was not significant: $F(4, 530) = .51$, $p = .73$, Partial $\eta^2 = .004$.

The Distinctive Feature x Warning interaction was not significant: $F(2, 530) = .36$, $p = .70$, Partial $\eta^2 = .001$.

The Number of Witness Groups x Warning interaction was not significant: $F(2, 530) = .60$, $p = .55$, Partial $\eta^2 = .002$.

The Number of Witness Groups x Warning x Distinctive Feature interaction was not significant: $F(4, 530) = 1.23$, $p = .29$, Partial $\eta^2 = .01$.

Supplementary Analyses for Experiment 3

Compliance Check Exclusions

We ran a 4 (Number of Witnesses: 5; 10; 15; 20) x 2 (Lineup Administration: biased; not biased) ANOVA, with guilt rating as the dependent variable, excluding the 14.7% who failed the compliance checks (see also Appendix B, for the compliance questions and the percentage of subjects who failed each one).

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 403) = .35$, $p = .79$, Partial $\eta^2 = .003$.

The lineup administration main effect was significant: $F(1, 403) = 68.36$, $p < .001$, Partial $\eta^2 = .15$. Subjects who were not told the lineup administration was biased gave higher guilt ratings ($M = 79.16$ $SD = 19.44$) than subjects who were told the lineup administration was biased ($M = 59.90$, $SD = 26.85$).

The number of witnesses main effect was not signficant: $F(3, 403) = .63$, $p = .60$, Partial $\eta^2 = .01$.

We ran a logistic regression with number of witnesses (5, 10, 15, 20) and verdict as the outcome variable, excluding the 14.7% who failed the compliance checks. The model was significant overall $x^2 (4) = 49.14$, $p < .001$. There was no evidence that the number of witnesses influenced guilty verdicts (Wald = 1.61, $p = .66$). The model showed people in the biased experiment were less likely to give guilty verdicts than the not biased experiment (Wald = 43.34, $p < .001$).

Attention Check Exclusions

We ran a 4 (Number of Witnesses: 5; 10; 15; 20) x 2 (Lineup Administration: biased; not biased) ANOVA, with guilt rating as the dependent variable, excluding the 16.6% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 394) = .27$, $p = .85$, Partial $\eta^2 = .002$.

The lineup administration main effect was significant: $F(1, 394) = 69.53$, $p < .001$, Partial $\eta^2 = .15$. Subjects who were not told the lineup administration was biased gave higher guilt ratings ($M = 80.87$, $SD = 19.32$) than subjects who were told the lineup administration was biased ($M = 61.46$, $SD = 26.72$).

The number of witnesses main effect was not significant: $F(3, 394) = .27$, $p = .85$, Partial $\eta^2 = .002$.

We ran a logistic regression with number of witnesses (5, 10, 15, 20) and verdict (guilty, not guilty) as the predictor variables and verdict as the outcome variable, excluding the 16.6% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup. The model was significant overall $x^2 (4) = 49.47$, $p < .001$. There was no evidence that the number of witnesses influenced guilty verdicts (Wald = .79, $p = .85$). The model showed people in the biased experiment were less likely to give guilty verdicts than the not biased experiment (Wald = 43.93, $p < .001$).

Lineup Fairness Ratings

We ran a 4 (Number of Witnesses: 5; 10; 15; 20) x 2 (Lineup Administration: biased; not biased) ANOVA, with lineup fairness ratings as the dependent variable. The significant main effect of bias was presented in Chapter 2 ($F(1, 474) = 99.29$ $p < .001$, Partial $\eta^2 = .17$). The remaining, non-significant, main effect and interaction are presented below.

The number of witnesses main effect was not significant: $F(3, 474) = .31$, $p = .82$, Partial $\eta^2 = .002$.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 474) = .82$, $p = .49$, Partial $\eta^2 = .005$.

Lineup Administration Fairness Ratings

We ran a 4 (Number of Witnesses: 5; 10; 15; 20) x 2 (Lineup Administration: biased; not biased) ANOVA, with lineup administration fairness ratings as the dependent variable. The significant main effect of bias was presented in Chapter 2 ($F(1, 474) = 157.25$ $p < .001$, Partial $\eta^2 = .25$). The remaining, non-significant, main effect and interaction are presented below.

The number of witnesses main effect was not significant: $F(3, 474) = .16$, $p = .92$, Partial $\eta^2 = .001$.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 474) = .34$, $p = .80$, Partial $\eta^2 = .002$.

Supplementary Analyses for Experiment 4

Primary Analysis of Verdict

A total of 29.2% of subjects in the 2 then 6 condition gave a guilty verdict for Case 1 and 34.0% in Case 2. We ran an exact McNemar's test which was not statistically significant ($p = .33$).

A total of 34.7% of subjects in the 6 then 2 condition gave a guilty verdict for Case 1 and 25.5% in Case 2. We ran an exact McNemar's test which was not statistically significant ($p = .06$).

Compliance Check Exclusions

We ran a paired-samples *t*-tests examining the first and second guilt ratings for the subjects in the 2 then 6 condition, excluding the 20.6% of subjects who failed the compliance checks, which was not significant $t(84) = 1.38$, $p = .17$, Cohen's $d = 0.09$.

We ran a paired-samples *t*-tests examining the first and second guilt ratings for the subjects in the 6 then 2 condition, excluding the 20.6% of subjects who failed the compliance checks which was not significant $t(76) = 1.64$, $p = .11$, Cohen's $d = 0.16$.

We ran a 2 (Order: 2 then 6; 6 then 2) x 2 (Guilt Rating: Case 1 rating; Case 2 rating) mixed ANOVA, with order as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 20.6% who failed the compliance checks.

The Guilt Rating x Order interaction was significant: $F(1, 160) = 4.66$, $p = .03$, Partial $\eta^2 = .03$.

The guilt rating main effect was not significant, $F(1, 160) = .13$, $p = .72$, Partial $\eta^2 = .001$.

The order main effect was not significant, $F(1, 160) = .84$, $p = .36$, Partial $\eta^2 = .01$

We then tested verdicts. A total of 27.1% of subjects in the 2 then 6 condition gave a guilty verdict for Case 1 and 35.3% in Case 2. We ran an exact McNemar's test which was not statistically significant ($p = .07$).

A total of 27.3% of subjects in the 6 then 2 condition gave a guilty verdict for Case 1 and 19.5% in Case 2. We ran an exact McNemar's test which was not statistically significant ($p = .21$).

Attention Check Exclusions

We ran a paired-samples *t*-test examining the Case 1 and Case 2 guilt ratings for the subjects in the 2 then 6 condition, excluding the 65.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup which was significant*:* $t(40) = 3.41$, $p = .002$, Cohen's $d = 0.28$.

We ran a paired-samples *t*-tests examining the Case 1 and Case 2 guilt ratings for the subjects in the 6 then 2 condition, excluding the 65.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup which was not significant $t(29) = 1.46$, $p = .16$, Cohen's $d = 0.24$.

We ran a 2 (Order: 2 then 6; 6 then 2) x 2 (Guilt Rating: Case 1 ratings; Case 2 ratings) mixed ANOVA, with order as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 65.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Guilt Rating x Order interaction was significant: $F(1, 69) = 10.42$, $p = .002$, Partial $\eta^2 = .13$.

The guilt rating main effect was not significant: $F(1, 69) = .53$, $p = .47$, Partial $\eta^2 = .01$.

The order main effect was not significant: $F(1, 69) = .00$, $p = .99$, Partial $\eta^2 = .000$.

We then examined verdict. A total of 17.1% of subjects in the 2 then 6 condition gave a guilty verdict for Case 1 and 29.3% in Case 2. We ran an exact McNemar's test which was not statistically significant ($p = .06$).

A total of 26.7% of subjects in the 6 then 2 condition gave a guilty verdict for Case 1 and 20.0% in Case 2 We ran an exact McNemar's test which was not statistically significant ($p = .63$).

Supplementary Analyses for Experiment 5

Primary Analysis of Verdict

We first examined within-subjects effects. A total of 63.2% of subjects in the 2 then 6 condition gave a guilty verdict for Time 1 and 88.7% for Time 2. We ran an exact McNemar's test which was statistically significant ($p < .001$).

A total of 81.8% of subjects in the 10 then 6 condition gave a guilty verdict for Time 1 and 71.7% for Time 2. We ran an exact McNemar's test which was statistically significant ($p = .02$).

We then examined between-subjects effects. To examine the between subjects-effects, we ran two logistic regressions, one with Verdict 1 as the dependent variable and one with Verdict 2 as the dependent variable. First, we compared the 63.2% of subjects who gave a guilty verdict when they were told two witnesses identified the suspect and the 81.8% of subjects who gave a guilty verdict when they were told 10 witnesses identified the suspect, which was significant x^2 (1) = 9.01, $p = .003$ (Wald = 8.55, $p = .003$). Second, we compared the 88.7% of subjects in the 2 then 6 condition who gave a guilty verdict when they were told six witnesses identified the suspect and 71.7% of subjects in the 10 then 6 condition who gave a guilty verdict when they were told six witnesses identified the suspect, which was significant x^2 (1) = 9.56, $p = .002$ (Wald = 8.85, $p = .003$).

Compliance Check Exclusions

We ran an independent samples *t*-test with number of witnesses as the independent variable and Time 1 guilt ratings as the dependent variable, excluding the 30.2% of subjects who failed the compliance checks which was significant*:* $t(136.91) = 5.87$, $p < .001$, Cohen's $d = 0.98$ (with correction for unequal variances). Subjects who were told 10

witnesses identified the suspect gave higher guilt ratings (M = 80.59, SD = 17.76) than subjects who were told 2 witnesses identified the suspect (M = 61.28, SD = 21.45).

We then ran a 2(Number of Witnesses: 10 then 6; 2 then 6) x 2(Guilt Rating: Time 1; Time 2) mixed ANOVA, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 30.2% of subjects who failed the compliance checks.

The Guilt Rating x Number of Witnesses interaction was significant: $F(1, 141) = 83.63$, $p < .001$, Partial $\eta^2 = .37$. To further interpret this interaction we ran two paired-samples t-tests. First, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 2 then 6 condition. The t-test was significant $t(71) = 9.02$, $p < .001$, Cohen's $d = 1.26$. Subjects gave significantly higher guilt ratings when they were told there were six witnesses (M = 87.25, SD = 19.63) compared to when they were told there were two witnesses (M = 61.28, SD = 21.45). Second, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 10 then 6 condition. The t-test was significant $t(70) = 2.71$, $p = .01$, Cohen's $d = 0.23$. Subjects who were first told 10 witnesses identified the suspect (M = 80.59, SD = 17.76) significantly adjusted their guilt ratings downwards after being told there was an error and actually 6 witnesses identified the suspect (M = 76.17, SD = 19.87).

We then examined differences in these adjustments. We ran an independent samples t-test with number of witnesses as the independent variable and second guilt rating as the dependent variable, $t(141) = 3.36$, $p = .001$, Cohen's $d = 0.56$. Guilt ratings were higher for subjects when they were told there were 6 witnesses for those in the 2 then 6 condition (M = 87.25, SD = 19.63) than the 10 then 6 condition (M = 76.17, SD = 19.87).

We then examined within-subjects effects of verdict. A total of 62.5% of subjects in the 2 then 6 condition gave a guilty verdict for the first case and 88.9% in the second case. We ran an exact McNemar's test which was statistically significant ($p < .001$).

A total of 84.5% of subjects in the 10 then 6 condition gave a guilty verdict for the first case and 76.1% in the second case. We ran an exact McNemar's test which was not statistically significant ($p = .07$).

We then examined between-subjects effects of verdict. To examine the between subjects-effects, we ran two logistic regressions, one with Verdict 1 as the dependent variable and one with Verdict 2 as the dependent variable. First, we compared the 62.5% of subjects who gave a guilty verdict when they were told two witnesses identified the suspect and the 84.5% of subjects who gave a guilty verdict when they were told 10 witnesses identified the suspect, which was significant $x^2 (1) = 9.01$, $p = .003$ (Wald = 8.43, $p = .004$). Second, we compared the 88.9% of subjects in the 2 then 6 condition who gave a guilty verdict when they were told six witnesses identified the suspect and 76.1% of subjects in the 10 then 6 condition who gave a guilty verdict when they were told six witnesses identified the suspect, which was significant $x^2 (1) = 4.16$, $p = .04$ (Wald = 3.91, $p = .05$).

Attention Check Exclusions

We ran an independent samples *t*-test with number of witnesses as the independent variable and Time 1 guilt ratings as the dependent variable, excluding the 33.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup. The *t*-test was significant: $t(135) = 5.21$, $p < .001$, Cohen's $d = 0.89$. Subjects who were told 10 witnesses identified the suspect gave higher guilt ratings ($M = 79.96$, $SD = 18.16$) than subjects who were told 2 witnesses identified the suspect ($M = 61.86$, $SD = 22.46$).

We then ran a 2 (Number of Witnesses: 10 then 6; 2 then 6) x 2 (Guilt Rating: time 1; time 2) mixed ANOVA, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 33.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Guilt Rating x Number of Witnesses interaction was significant: $F(1, 135) = 137.57$, $p < .001$, Partial $\eta^2 = .51$. To further interpret this interaction, we ran two paired-samples *t*-tests. First, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 2 then 6 condition. The *t*-test was significant: $t(64) = 11.99$, $p < .001$, Cohen's $d = 1.36$. Subjects gave significantly higher guilt ratings when they were told there were six witnesses ($M = 89.84$, $SD = 18.37$) compared to when they were told there were two witnesses ($M = 61.86$, $SD = 22.46$). Second, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 10 then 6 condition. The *t*-test was significant: $t(71) = 2.50$, $p = .01$, Cohen's $d = 0.19$. Subjects who were first told 10 witnesses identified the suspect ($M = 79.96$, $SD = 18.16$) significantly adjusted their guilt ratings downwards after being

told there was an error and actually 6 witnesses identified the suspect (M = 76.32, SD = 19.75).

We then examined differences in these adjustments. We ran an independent samples t-test with number of witnesses as the independent variable and second guilt rating as the dependent variable, $t(135) = 4.14$, $p < .001$, Cohen's $d = 0.71$. Guilt ratings were higher for subjects when they were told there were 6 witnesses for those in the 2 then 6 condition (M = 89.85, SD = 18.37) than the 10 then 6 condition (M = 76.32, SD = 19.75).

We then examined within-subjects effects of verdict. A total of 58.5% of subjects in the 2 then 6 condition gave a guilty verdict for the first case and 89.2% in the second case. We ran an exact McNemar's test which was statistically significant ($p < .001$).

A total of 81.9% of subjects in the 10 then 6 condition gave a guilty verdict for the first case and 76.4% in the second case. We ran an exact McNemar's test which was not statistically significant ($p = .29$).

We then examined between-subjects effects of verdict. To examine the between subjects-effects, we ran two logistic regressions, one with Verdict 1 as the dependent variable and one with Verdict 2 as the dependent variable. First, we compared the 58.5% of subjects who gave a guilty verdict when they were told two witnesses identified the suspect and the 81.9% of subjects who gave a guilty verdict when they were told 10 witnesses identified the suspect, which was significant $x^2 (1) = 9.23$, $p = .002$ (Wald = 8.72, $p = .003$). Second, we compared the 89.2% of subjects in the 2 then 6 condition who gave a guilty verdict when they were told six witnesses identified the suspect and 76.4% of subjects in the 10 then 6 condition who gave a guilty verdict when they were told six

witnesses identified the suspect, which was significant x^2 (1) = 4.02, p = .05 (Wald = 3.73, p = .05).

Supplementary Analyses for Experiment 6

Primary Analysis

Verdict. We analysed verdict as the dependent variable. We ran a logistic regression with number of witnesses (1, 2, 3, 4) and lineup administration (biased, not biased) as the predictor variables and verdict (guilty, not guilty) as the outcome variable. The model was not significant overall $\boldsymbol{x}^2(4) = 4.91$, p = .30 There was no evidence that the number of witnesses influenced guilty verdicts (Wald = 4.82, p = .19). There was also no evidence that people in the biased experiment were less likely to give guilty verdicts than the not biased experiment (Wald = .01, p = .93). Both of these findings are inconsistent with the significant administration and biased main effects found using our ANOVA model.

Compliance Check Exclusions

We ran a 4 (Number of Witnesses: 1; 2; 3; 4) x 2 (Lineup Administration: biased; not biased) ANOVA, with guilt rating as the dependent variable, excluding the 41.0% who failed the compliance checks (see also Appendix x, for the compliance questions and the percentage of subjects who failed each one).

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 235) = 1.11$, p = .35, Partial η^2 = .01.

The lineup administration main effect was significant: $F(1, 235) = 18.04$, $p < .001$, Partial η^2 = .07. Subjects who were not told the lineup administration was biased gave

higher guilt ratings ($M = 69.12$, $SD = 23.97$) than subjects who were told the lineup administration was biased ($M = 55.45$, $SD = 25.32$).

The number of witnesses main effect was significant: $F(3, 235) = 9.40$, $p < .001$, Partial $\eta^2 = .12$. To further examine this main effect, we ran Tukey post-hoc tests. Subjects who were told one witness identified the suspect gave lower guilt ratings than those subjects who were told two witnesses ($p = .003$), three witnesses ($p < .001$), and four witnesses ($p < .001$) identified the suspect. There were no other significant differences.

We ran a logistic regression with number of witnesses (1, 2, 3, 4) and verdict as the outcome variable, excluding the 38.3% who failed the compliance checks. The model was significant overall $x^2 (4) = 15.80$, $p = .003$. The number of witnesses increased guilty verdicts (Wald = 12.65, $p = .01$). The model showed no evidence that people in the biased experiment were less likely to give guilty verdicts than the not biased experiment (Wald = 2.29, $p = .13$).

Attention Check Exclusions

We ran a 4 (Number of Witnesses: 1; 2; 3; 4) x 2 (Lineup Administration: biased; not biased) ANOVA, with guilt rating as the dependent variable, excluding the 38.3% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 246) = 1.25$, $p = .29$, Partial $\eta^2 = .02$.

The lineup administration main effect was significant: $F(1, 246) = 18.63$, $p < .001$, Partial $\eta^2 = .07$. Subjects who were not told the lineup administration was biased gave higher guilt ratings ($M = 69.22$, $SD = 24.25$) than subjects who were told the lineup administration was biased ($M = 54.91$, $SD = 24.60$).

The number of witnesses main effect was significant: $F(3, 246) = 8.41$, $p < .001$, Partial $\eta^2 = .09$. To further examine this main effect, we ran Tukey post-hoc tests. Subjects who were told one witness identified the suspect gave lower guilt ratings than those subjects who were told two witnesses ($p = .004$), three witnesses ($p < .001$), and four witnesses ($p < .001$) identified the suspect. There were no other significant differences.

We ran a logistic regression with number of witnesses (1, 2, 3, 4) and verdict (guilty, not guilty) as the predictor variables and verdict as the outcome variable, excluding the 38.3% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup. The model was significant overall $x^2 (4) = 14.75$, $p = .01$. The number of witnesses increased guilty verdicts (Wald = 11.87, $p = .01$). There was no evidence that

people in the biased experiment were less likely to give guilty verdicts than the not biased experiment (Wald = 1.84, $p = .18$).

Lineup Fairness Ratings

We ran a 4 (Number of Witnesses: 1; 2; 3; 4) x 2 (Lineup Administration: biased; not biased) ANOVA, with lineup fairness ratings as the dependent variable. The significant main effect of bias was presented in Chapter 2 ($F(1, 16) = 24.03$ $p < .001$, Partial $\eta^2 = .06$). The remaining, non-significant, main effect and interaction are presented below.

The number of witnesses main effect was not significant: $F(3, 416) = .78$, $p = .50$, Partial $\eta^2 = .01$.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 416) = 1.60$, $p = .19$, Partial $\eta^2 = .01$.

Lineup Administration Fairness Ratings

We ran a 4 (Number of Witnesses: 1; 2; 3; 4) x 2 (Lineup Administration: biased; not biased) ANOVA, with lineup administration fairness ratings as the dependent variable. The significant main effect of bias was presented in Chapter 2 ($F(1, 411) = 36.77$ $p < .001$, Partial $\eta^2 = .08$). The remaining, non-significant, main effect and interaction are presented below.

The number of witnesses main effect was not significant: $F(3, 411) = .83$, $p = .48$, Partial $\eta^2 = .01$.

The Number of Witnesses x Lineup Administration interaction was not significant: $F(3, 411) = 1.64$, $p = .18$, Partial $\eta^2 = .01$.

Supplementary Analyses for Experiment 7

Primary Analysis of Verdict

We examined within-subjects effects. A total of 82.6% of subjects in the 5 then 10 condition gave a guilty verdict for the first case and 87.2% in the second case. We ran an exact McNemar's test which was not statistically significant ($p < .13$).

A total of 85.8% of subjects in the 15 then 10 condition gave a guilty verdict for the first case and 77.4% in the second case. We ran an exact McNemar's test which was statistically significant ($p = .01$).

We then examined between-subjects effects. To examine the between subjects-effects, we ran two logistic regressions, one with Verdict 1 as the dependent variable and one with Verdict 2 as the dependent variable. First, we compared the 82.6% of subjects who gave a guilty verdict when they were told five witnesses identified the suspect and the 85.8% of subjects who gave a guilty verdict when they were told 15 witnesses identified the suspect, which was not significant x^2 (1) = .44, p = .51 (Wald = .43, p = .51). Second, we compared the 87.2% of subjects in the 5 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect and 77.4% of subjects in the 15 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect, which was not significant x^2 (1) = 3.58, p = .06 (Wald = 3.47, p = .06).

Compliance Check Exclusions

We ran an independent samples *t*-test with number of witnesses as the independent variable and Time 1 guilt ratings as the dependent variable, excluding the 27.0% of subjects who failed the compliance checks. The *t*-test was not significant: $t(155) = 1.70$, p = .09, Cohen's d = 0.27.

We then ran a 2 (Number of Witnesses: 10 then 6; 2 then 6) x 2 (Guilt Rating: Time 1; Time 2) mixed ANOVA, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 27.0% of subjects who failed the compliance checks.

The Guilt Rating x Number of Witnesses interaction was significant: $F(1, 155) = 40.90$, $p < .001$, Partial $\eta^2 = .21$. To further interpret this interaction, we ran two paired-samples *t*-tests. First, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 5 then 10 condition. The *t*-test was significant $t(81) = 6.74$, $p < .001$, Cohen's $d = 0.66$. Subjects gave significantly higher guilt ratings when they were told there were ten witnesses ($M = 90.51$, $SD = 14.98$) compared to when they were told there were five witnesses ($M = 79.67$, $SD = 17.86$). Second, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 15 then 10 condition. The *t*-test was significant $t(74) = 2.12$, $p = .04$, Cohen's $d = 0.16$. Subjects who were first told 15 witnesses identified the suspect ($M = 84.41$, $SD = 17.13$) significantly adjusted their guilt ratings downwards after being told there was an error and actually 10 witnesses identified the suspect ($M = 81.39$, $SD = 20.45$).

We then examined differences in adjustments. We ran an independent samples *t*-test with number of witnesses as the independent variable and second guilt rating as the dependent variable, $t(155) = 3.21$, $p = .002$, Cohen's $d = 0.51$. Guilt ratings were higher for subjects when they were told there were 10 witnesses for those in the 5 then 10 condition ($M = 90.51$, $SD = 14.98$) than the 15 then 10 condition ($M = 81.39$, $SD = 20.45$).

We then examined within-subjects effects of verdict. A total of 85.4% of subjects in the 5 then 10 condition gave a guilty verdict for the first case and 87.8% in the second case. We ran an exact McNemar's test which was not statistically significant ($p = .63$).

A total of 84.0% of subjects in the 15 then 10 condition gave a guilty verdict for the first case and 76.0% in the second case. We ran an exact McNemar's test which was statistically significant ($p = .03$).

We then examined between-subjects effects of verdict. To examine the between subjects-effects, we ran two logistic regressions, one with verdict 1 as the dependent variable and one with verdict 2 as the dependent variable. First, we compared the 85.4% of subjects who gave a guilty verdict when they were told five witnesses identified the suspect and the 84.0% of subjects who gave a guilty verdict when they were told 15 witnesses identified the suspect, which was significant $x^2(1) = .56$, $p = .81$ (Wald = .06, $p = .81$). Second, we compared the 87.8% of subjects in the 5 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect and 76.0% of subjects in the 15 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect, which was significant, with an ambiguous p-value $x^2(1) = 3.75$, $p = .05$ (Wald = 3.61, $p = .06$).

Attention Check Exclusions

We ran an independent samples *t*-test with number of witnesses as the independent variable and Time 1 guilt ratings as the dependent variable, excluding the 27.0% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup. The t-test was not significant*:* $t(155) = 1.49$, $p = .14$, Cohen's $d = 0.24$.

We then ran a 2 (Number of Witnesses: 10 then 6; 2 then 6) x 2 (Guilt Rating: Time 1; Time 2) mixed ANOVA, with number of witnesses as the between-subjects independent variable and guilt rating as the repeated measure, excluding the 33.2% who could not recall at least one of the two questions pertaining to how many witnesses they were told saw the crime and how many witnesses identified the suspect from the lineup.

The Guilt Rating x Number of Witnesses interaction was significant: $F(1, 155) = 56.16$, $p < .001$, Partial $\eta^2 = .27$. To further interpret this interaction, we ran two paired-samples *t*-tests. First, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 5 then 10 condition. The *t*-test was significant $t(78) = 8.15$, $p < .001$, Cohen's $d = 0.73$. Subjects gave significantly higher guilt ratings when they were told there were ten witnesses ($M = 92.42$, $SD = 13.02$) compared to when they were told there were five witnesses ($M = 80.32$, $SD = 19.62$). Second, we examined the Time 1 and Time 2 guilt ratings for the subjects in the 15 then 10 condition. The *t*-test was not significant $t(77) = 1.95$, $p = .06$, Cohen's $d = 0.13$.

We then examined differences in adjustments. We ran an independent samples *t*-test with number of witnesses as the independent variable and Time 2 guilt rating as the dependent variable, $t(155) = 3.91$, $p < .001$, Cohen's $d = 0.62$. Guilt ratings were higher for

subjects when they were told there were 10 witnesses for those in the 5 then 10 condition (M = 92.42, SD = 13.02) than the 15 then 10 condition (M = 82.29, SD = 18.91).

We then examined within-subjects effects of verdict**.** A total of 86.1% of subjects in the 5 then 10 condition gave a guilty verdict at Time 1 and 88.6% at Time 2. We ran an exact McNemar's test which was not statistically significant ($p < .50$).

A total of 83.3% of subjects in the 15 then 10 condition gave a guilty verdict at Time 1 and 78.2% at Time 2. We ran an exact McNemar's test which was not statistically significant ($p = .22$).

We then examined between-subjects effects of verdict. To examine the between subjects-effects, we ran two logistic regressions, one with Verdict 1 as the dependent variable and one with Verdict 2 as the dependent variable. First, we compared the 86.1% of subjects who gave a guilty verdict when they were told five witnesses identified the suspect and the 83.3% of subjects who gave a guilty verdict when they were told 15 witnesses identified the suspect, which was not significant x^2 (1) = .23, p = .63 (Wald = .23, p = .63). Second, we compared the 88.6% of subjects in the 5 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect and 78.2% of subjects in the 15 then 10 condition who gave a guilty verdict when they were told ten witnesses identified the suspect, which was significant x^2 (1) = 3.11, p = .08 (Wald = 2.98, p = .08).

Appendix D

Table D1

Means and Standard Deviations for each Condition in Experiment 1

		2 then 12		10 then 12	
		Time 1	Time 2	Time 1	Time 2
Left Scar	Warned	59.16 (23.87)	75.78 (21.58)	57.45 (20.10)	67.47 (22.83)
	Not warned	72.65 (21.61)	86.83 (19.66)	80.4 (20.69)	87.54 (20.13)
Right Scar	Warned	57.96 (19.25)	73.20 (23.36)	58.08 (23.36)	66.73 (23.53)
	Not warned	76.74 (17.10)	90.40 (15.13)	79.19 (22.38)	83.11 (23.92)
Black eye	Warned	54.69 (25.58)	67.47 (25.92)	56.43 (24.88)	74.00 (26.59)
	Not warned	74.31 (20.57)	85.96 (22.07)	69.06 (24.19)	81.33 (23.13)

Note. Standard deviations are presented in parentheses.

Table D2

Means and Standard Deviations for each Condition in Experiment 2

		1 group	2 groups		3 groups	
		-	After each group	At the end	After each group	At the end
Left Scar	Warned	60.88 (28.74)	75.37 (18.02)	68.25 (20.05)	73.14 (30.56)	68.84 (25.08)
	Not warned	80.00 (24.36)	89.42 (13.78)	83.89 (18.81)	83.89 (20.55)	82.00 (22.57)
Right Scar	Warned	58.67 (22.68)	70.75 (26.26)	77.20 (17.65)	67.71 (28.83)	69.90 (17.11)
	Not warned	78.65 (20.15)	88.20 (19.87)	84.88 (14.28)		86.86 (16.76)
Black eye	Warned	64.56 (25.45)	62.58 (26.25)	63.11 (28.180	71.38 (30.31)	74.14 (27.21)
	Not warned	74.95 (21.42)	86.13 (12.96)	85.50 (16.34)	75.36 (31.09)	85.86 (12.36)

Note. Standard deviations are presented in parentheses. Due to an error, there were no subjects assigned to the right scar, three witness group, not warned condition.

Table D3

Means and Standard Deviations for each Condition in Experiment 3

	5	10	15	20
Biased	59.43 (27.49)	62.00 (26.26)	62.07 (25.48)	59.49 (27.04)
Not biased	78.10 (16.97)	78.82 (22.12)	81.64 (17.24)	79.44 (22.98)

Note. Standard deviations are presented in parentheses.

Table D4

Means and Standard Deviations for each Condition in Experiment 4

	Case 1	Case 2
2 then 6	43.08 (24.61)	45.33 (27.41)
6 then 2	44.51 (22.42)	40.26 (21.23)

Note. Standard deviations are presented in parentheses.

Table D5

Means and Standard Deviations for each Condition in Experiment 5

	Time 1	Time 2
2 then 6	61.81 (22.06)	83.24 (21.76)
10 then 6	77.64 (19.16)	74.56 (20.09)

Note. Standard deviations are presented in parentheses.

Table D6

Means and Standard Deviations for each Condition in Experiment 6

	1	2	3	4
Biased	51.45 (22.58)	65.15 (19.42)	65.44 (22.88)	58.65 (22.39)
Not biased	58.63 (23.36)	66.72 (18.84)	75.35 (19.90)	72.16 (26.64)

Note. Standard deviations are presented in parentheses.

Table D7

Means and Standard Deviations for each Condition in Experiment 7

	Time 1	Time 2
5 then 10	77.21 (20.59)	87.38 (18.15)
15 then 10	81.09 (18.91)	80.15 (19.91)

Note. Standard deviations are presented in parentheses.

Appendix E

CONFIDENTIAL - FOR PEER-REVIEW ONLY

ASPREDICTED

Robbie Taylor Thesis E6a (#15769)

Created: 10/30/2018 04:06 PM (PT)
Shared: 04/12/2020 03:20 PM (PT)

This pre-registration is not yet public. This anonymized copy (without author names) was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) will become publicly available only if an author makes it public. Until that happens the contents of this pre-registration are confidential.

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does the number of witnesses who unanimously pick a suspect from a lineup influence jurors guilt ratings?

3) Describe the key dependent variable(s) specifying how they will be measured.
Guilt ratings (0-100)
Verdict (Guilty versus not guilty)

4) How many and which conditions will participants be assigned to?
IV: number of witnesses (1, 2, 3, or 4)

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Primary analysis: A one-way ANOVA to compare between the 4 conditions on the guilt ratings

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude people who do not finish. We MAY exclude people who fail out attention checks (For example, did you maximise your screen?) if the results of those who failed change the overall pattern of results. We may also exclude those who fail our manipulation checks (How many witnesses identified person x?) if including or excluding these people changes the pattern of findings or significance of our tests. In that case, we will report all findings twice: first, the analysis including these subjects, and second, the analysis excluding these subjects

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
200. 50 per cell. This is in line with our previous experiments

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.

ASPREDICTED

Robbie's Thesis Experiment 6B (#15775)

Created: 10/30/2018 08:50 PM (PT)
Shared: 04/12/2020 03:20 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
To what extent does the number of unanimous witnesses who identify a suspect from a lineup influences jurors' ratings of guilt?
To what extent does this relationship change when the lineup administration is biased versus not biased?

Based on our previous experiments we might expect to see no difference between the number of witnesses and guilt ratings.
In line with the conformity literature, we might see that after the addition of 3 witnesses, the relationship does not further increase.

We expect a main effect versus biased and non-biased lineup administration. The biased condition should have lower guilt ratings

3) Describe the key dependent variable(s) specifying how they will be measured.
Guilt ratings (0-100)
Verdict (Guilty versus not guilty)

4) How many and which conditions will participants be assigned to?
Number of witnesses (1, 2, 3, or 4)
Unlike Experiment 6A, the lineups in these conditions are not biased—that is the only difference between E6A and E6B

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Primary analyses: One way ANOVA comparing the 1, 2, 3, 4 conditions with respect to guilt ratings

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
The same as Experiment 6a:
We will exclude people who don't finish
We will see if excluding based on failing attention checks and manipulation checks influence the pattern of findings. If they do, we will present the findings twice: first, including these subjects; second, excluding these subjects.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
200. 50 per cell. The same as our previous experiments

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.

ASPREDICTED

Unanimous Witnesses Experiment 7. (#18990)

Created: 01/27/2019 01:39 PM (PT)
Shared: 04/12/2020 03:21 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We have two non-mutually exclusive explanations for one of previous studies: 1) After a group size of 3 witnesses, there is no evidence to suggest any additional witnesses coming forward continues to increase guilt ratings 2) Subjects believe the suspect is probably guilty, so upon learning about evidence that confirms that belief, they adjust their guilt ratings upwards. But upon learning about evidence that is counter to their belief, they do not adjust their guilt ratings downwards—at least to the same extent as they would adjust upwards. Therefore, if we redo Experiment 5, but use higher numbers of witnesses, so all witness are above the number 3, we can remove the influence of the first explanation. And therefore the differences observed are likely to be caused by confirmation bias. We suspect we will observe a similar pattern as Experiment 5. Although, we would expect the difference between the 5 and 10 condition will not be as large as the difference between the 2 and 6 condition in experiment 5 If the size of the group does matter beyond 3—which we have seen mixed evidence for in our previous experiments—then we might expect overall guilt ratings to be higher in the current experiment than Experiment 5

3) Describe the key dependent variable(s) specifying how they will be measured.
Guilt ratings Guilt verdicts

4) How many and which conditions will participants be assigned to?
two conditions: 5 then 10 condition and 15 then 10 condition

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Between subjects t-tests or one way ANOVAs for: First rating comparing 5 and 15 witnesses—we have seen mixed evidence for a between-subjects difference in guilt ratings, but we might expect to see the 15 group is slightly higher than the 5 on guilt ratings Second ratings comparing both groups when they are presented with 10 witnesses. Like Experiment 5, we expect the 15 then 10 group to give higher second ratings than the 5 then 15

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude people who do not finish If including people who failed the attention checks and manipulation checks change the overall pattern of results, we will report the main findings both with and without those subjects included

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
200. 100 per cell

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing to add