University of Nevada, Reno

Effects of Evidence Strength, Format, and Order on Police Judgments of Suspect Guilt

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy in

Social Psychology

by

Jean J. Cabell

Dr. Yueran Yang/Dissertation Advisor

May, 2022

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

JEAN J. CABELL

entitled

Effects of Evidence Strength, Format, and Order on Police Judgments of Suspect Guilt

be accepted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Yueran Yang, Ph.D.
Advisor

Emily R. Berthelot, Ph.D.
Committee Member

Shawn C. Marsh, Ph.D.
Committee Member

Monica K. Miller, Ph.D.
Committee Member

Markus K. Kemmelmeier, Ph.D.
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

May, 2022

Abstract

Police encounter evidence when conducting criminal investigations. This evidence informs their judgments of suspect guilt and can affect investigative decisions. Yet, surprisingly little research examines how evidence affects police judgments of suspect guilt. This dissertation contains three papers to address this gap. Paper #1 is a theoretical review paper that proposed a unified definition of evidence strength based on competing hypotheses, as varying definitions of evidence strength could lead to inaccuracy when police use evidence to judge suspect guilt. Paper #2 is an empirical paper that recruited police officers (N = 209) to examine the effects of evidence strength format and evidence type on police guilt judgments. I assigned participants to a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-subjects factorial design and asked them to judge suspect guilt on two measures. Overall, participants were most accurate when encountering evidence strength in an LR format but only for DNA evidence. Paper #3 is an empirical paper that recruited police (N = 75) and laypeople (N = 636) to examine the effects of evidence order and social norms on police evaluations of evidence and judgments of suspect guilt. I randomly assigned participants to a 2 (evidence order: incriminating evidence first vs. incriminating evidence last) x 2 (social norms: efficiency vs. thoroughness) x 2 (type: DNA incriminating-eyewitness ambiguous vs. eyewitness incriminating-DNA ambiguous) between-subjects factorial design and asked them to evaluate the evidence and judge suspect guilt. Overall, social norms that prioritized a thorough investigation (vs. efficient) minimized biased evidence evaluations and guilt judgments. Thus, this dissertation expands upon previous psychology and law research to better understand

how evidence influences police guilt judgments.

Acknowledgments

Many people have supported me through my dissertation process who I would like to thank and acknowledge.

First and foremost, I would like to express immense gratitude to the Russell J. and Dorothy S. Bilinski Fellowship Fund, a program of the Bilinski Educational Foundation. This fellowship provided me with the financial means and invaluable flexibility to complete this project.

Thank you to my advisor, Dr. Yueran Yang, who helped with the ideation and structure of my theoretical paper and provided feedback during iterations of my experimental designs. I would also like to thank each member of my dissertation committee: Drs. Emily Berthelot, Markus Kemmelmeier, Shawn Marsh, and Monica Miller. I appreciated your support, encouragement, and enthusiasm throughout this project.

Thank you to everyone who helped me recruit police officers to participate in my dissertation research.

Thank you to the people graduate school brought into my life including but not limited to Tyler, Peter, Leah, Jonathan, and Caroline. You were each a vital component of my support system throughout this project.

Last, but certainly not least, I would like to express my deepest gratitude to my parents John and Lori Cabell. Thank you for your endless support throughout my educational pursuits.

Table of Contents

List of Tables

Chapter 3

Chapter 4

List of Figures

Chapter 1: Introduction

On June 13th, 1996, an 18-year-old woman named Angie Dodge was raped and murdered in her apartment. Police investigators identified Christopher Tapp as a potential suspect and coerced him to confess to the crimes (Murphy, 2019). DNA evidence collected from the crime scene did not support the hypothesis that Tapp committed the crime, but police investigators ignored this DNA evidence and maintained that Tapp was guilty. Tapp was convicted of these crimes in 1998. It was not until 2019 that analysts working with the Idaho Innocence Project identified another man, Brian Dripps, as a possible suspect for the crimes against Angie Dodge (Otterbourg, 2021). Dripps's DNA matched the DNA from the crime scene and Tapp was proven to be innocent of the crimes against Angie Dodge. Tapp was exonerated in 2019 after over 20 years in prison for a crime he did not commit. On February 9th, 2021, Brian Dripps pleaded guilty to first-degree murder and rape for the crime against Angie Dodge that occurred almost 25 years ago. This case is an example of failure by police investigators to accurately evaluate evidence when judging a suspect's guilt, and of how this inaccuracy led to the conviction and imprisonment of an innocent man.

Police investigators must judge the likelihood of a suspect's guilt, perhaps even multiple times, to make consequential investigative decisions. For example, police investigators might decide whether to gather more evidence, when to interrogate a suspect, whether to arrest a suspect, or when to refer a case to a prosecutor. However, police investigations could go awry if the police misidentify innocent people, such as Christopher Tapp, as suspects because police's mistaken judgments could lead them to make decisions that end in wrongful convictions (Scherr et al., 2020). Although these

consequential investigative decisions rely upon police properly evaluating evidence, surprisingly little research has examined how police evaluate evidence to inform their judgments of a suspect's guilt. To improve investigative practices and prevent wrongful convictions, it is therefore important to identify how police evaluate evidence to form their guilt judgments.

This dissertation contains three papers that aim to achieve three main objectives to further understand how police investigators evaluate evidence to improve investigative practices and ultimately minimize wrongful convictions. First, Paper #1 (*Evaluating evidence in a criminal investigation: Toward a unified definition of evidence strength*) proposed a unified operational definition of evidence strength for psychological research and police investigations. Second, Paper #2 (*Police comprehension of evidence: Statistical formats and evidence type*) examined police accuracy in incorporating a single piece of evidence into their guilt judgments and whether these judgments differed by the format or type of evidence. Third, Paper #3 (*Decreasing biased guilt judgments during an investigation with social norms*) examined how police and laypeople judged suspect guilt when there were multiple pieces of evidence by manipulating the order of the evidence and social norms during an investigation. This introduction chapter will provide a statement of the problem and the problem's importance before providing an overview of the theoretical foundation and methodology of each paper.

Problem Statement

Police collect evidence that can prove or substantiate facts related to a crime, such as corroborating suspect guilt or proving a crime occurred. Criminal evidence is critical for administering criminal justice during police investigations (Dror, 2018). Thus, it is

important to examine how police evaluate evidence because these evaluations could influence subsequent judgments of suspect guilt and decisions.

Police could error in judging suspect guilt if they do not properly evaluate evidence. However, several major gaps exist in addressing this problem. First, there is a lack of consensus within psychological literature on a consistent operationalization of evidence strength. Inconsistently defining evidence strength could lead police to inaccurately evaluate evidence if they are basing their decisions on research grounded in inconsistent definitions.

Second, there is a need to identify factors that influence how police evaluate evidence to judge suspect guilt during an investigation. The second two papers aim to address this gap. The goal of Paper #2 is to understand how the format to convey evidence strength and the type of evidence influence police accuracy in judging suspect guilt. The goal of Paper #3 is to better understand police evaluations of evidence and suspect guilt when they encounter evidence in different orders and when there are different social norms.

By addressing these problems, this dissertation provides intellectual merit through an interdisciplinary approach using concepts from social psychology, cognitive psychology, and statistics to meaningfully enrich the field of psychology and law in seven major ways: (1) by proposing a definition of evidence strength, (2) by distinguishing between objective and subjective evidence strength, (3) by replicating and extending findings on how mock jurors evaluate statistical classifications of evidence to the context of police investigations, (4) by examining how police evaluate statistical classifications of eyewitness identification evidence, (5) by clarifying mixed findings as

to whether police judgments of evidence and guilt align with confirmation bias, recency bias, or the Bayesian Cognitive Model when evaluating multiple pieces of evidence, (6) by examining a method to minimize bias induced by the order of evidence using social norms, (7) by recruiting police officer participants. Thus, the proposed dissertation furthers social psychology, cognitive psychology, and statistics within the psychology and law field.

Addressing these problems also has several broader impacts for a more just society to ultimately minimize wrongful convictions in four major ways: (1) by identifying a unified definition of evidence strength to improve consistency within police investigations, (2) by identifying the extent to which investigators are accurate in evaluating statistical classifications of a single piece of evidence, which has implications for police training, (3) by identifying a potential cognitive bias when police evaluate multiple pieces of evidence and proposing a method to minimize bias using social norms within a police department, which has implications for preventing investigative error and wrongful convictions, (4) by planning to disseminate findings from both studies through several means, including conference presentations and manuscript submissions.

Aims and Overview of Each Paper

Paper #1: Evaluating Evidence in a Criminal Investigation: Toward a Unified Definition of Evidence Strength

Existing psychological literature offers inconsistent operational definitions of evidence strength. Existing definitions of evidence strength include *the extent to which a piece of evidence is incriminating* (e.g., Lidén et al., 2019) or *the extent to which the evidence supports one hypothesis versus another hypothesis* (e.g., Robertson et al., 2016).

Some researchers are proponents of defining evidence strength based on the latter definition because it can precisely communicate evidence strength while also encompassing many different types of evidence. Although people might imply an alternative hypothesis in the former definition, it is not clear what the alternative hypothesis is when it is left unstated. Therefore, the former definition is not a precise way to operationalize evidence strength.

There is also a lack of nuance between subjective evidence strength (how people evaluate evidence) and objective evidence strength (how strong the evidence is in supporting any given proposition). This distinction is important because subjective evidence strength and objective evidence strength accomplish different research goals. Subjective evidence strength is important if the goal is to better understand how police evaluate evidence. Conversely, objective evidence strength is important because it provides a natural standard to assess accuracy if the goal is to understand the extent to which police are *accurate* in evaluating evidence.

Forensic experts use likelihood ratios (LRs) to communicate objective evidence strength (e.g., Robertson et al., 2016) and a growing body of research suggests using LRs when possible for various types of evidence (e.g., Horgan et al., 2012; Robertson et al., 2016; Steblay et al., 2011). I discuss the implications of a unified operationalization of evidence strength and of providing nuance between subjective and objective evidence strength in this paper.

Thus, this paper attempts to address the following gaps:

1) There are inconsistent operationalizations of evidence strength (e.g., Lidén et al., 2019; Robertson et al., 2016)

2) There is a need to separately define the subjective evidence strength from objective evidence strength to be able to assess police's accuracy in evaluating evidence

Chapter #2 contains the manuscript for Paper #1, which is a theoretical review and recommendations paper. This paper reviewed previous definitions of evidence strength, proposed a definition of evidence strength for research and practice, distinguished between subjective and objective evidence strength, proposed using likelihood ratios and logarithm likelihood ratios to clarify objective evidence strength, and provides recommendations for future research. The suggested publication outlet for this manuscript is *Applied Cognitive Psychology*. If publishing at *Applied Cognitive Psychology* is unsuccessful, another suggested publication outlet is *Law, Probability, and Risk*. I quantify my percentage effort for this article as 90% through conceptualizing the idea and writing the manuscript. The remaining 10% effort is attributed to my co-author and advisor, Yueran Yang, for helping with conceptualizing the idea and providing feedback on the written manuscript.

Paper #2: Police Comprehension of Evidence: Statistical Formats and Evidence Type

Two factors related to how evidence evaluations can influence police guilt judgments are how forensic experts communicate evidence strength (i.e., presentation format) and the type of evidence. Forensic experts and eyewitness researchers sometimes use an LR as the presentation format to convey evidence strength (e.g., Association of Forensic Science Providers, 2009; Thompson & Newman, 2015; Wells & Lindsay, 1980). LRs convey the probability that one hypothesis is true when compared to another

hypothesis. For example, an LR of 10 for fingerprint evidence conveys that it is 10 times more likely the two fingerprints originated from the same person than from different people. Eyewitness researchers also use LRs to convey how much more likely an eyewitness was to have identified the suspect if the suspect is guilty than if the suspect was innocent given that the eyewitness identified the suspect (e.g., Wells & Lindsay, 1980). The higher the LR, the stronger the evidence supports the proposition that the suspect is the culprit.

In addition to LRs, forensic experts also use a random match probability (RMP) to convey evidence strength. An RMP conveys the same information as an LR but in a frequency format. For example, an RMP of 10 for fingerprint evidence conveys that one person in 10 would have a fingerprint that is consistent with the fingerprint sample found at the crime scene. People tend to interpret frequencies more accurately than ratios (Gigerenzer & Hoffrage, 1995), so investigators should interpret evidence presented as an RMP more accurately than evidence presented as an LR. Mock jurors tend to evaluate DNA evidence accurately regardless of presentation format, but they were more accurate in evaluating shoeprint evidence in the RMP format than in an LR format (Thompson & Newman, 2015). However, Thompson and colleagues (2018) found some participants misunderstood RMPs. Therefore, people do not always interpret RMPs more accurately than LRs.

The type of evidence could also affect police accuracy in evaluating evidence. People tend to perceive DNA evidence as credible (Kassin et al., 2013; Lieberman et al., 2008; Thompson & Newman, 2015), and mock jurors tend to interpret DNA evidence accurately (Thompson & Newman, 2015). However, it is unknown how accurately police

(1) interpret DNA evidence, (2) evaluate DNA evidence compared to other types of forensic evidence, such as fingerprint evidence, (3) evaluate forensic evidence when compared to other types of evidence, such as eyewitness evidence. This knowledge is necessary to know which types of evidence on which police investigators need further training and which types of evidence are most prone to error during an investigation, particularly because evidence evaluations can vary depending on whether the evidence is within the context of a trial or an investigation (Sommers & Douglass, 2007). There is a lack of research that compares subjective evidence strength to an objective statistical standard of evidence strength for eyewitness identification (ID) evidence. Eyewitness ID evidence is one of the most persuasive and powerful types of evidence (e.g., Wells et al., 2006), and therefore it is critical to examine police accuracy in evaluating eyewitness ID evidence when judging a suspect's guilt.

Thus, this paper attempts to address the following gaps:

3) It is unknown how different formats of evidence strength (i.e., LRs and RMPs) influence the extent to which police are accurate in evaluating evidence to judge a suspect's guilt because previous research assessed evidence evaluations within the context of a criminal trial (Thompson et al., 2018; Thompson & Newman, 2015) and evaluations of evidence can differ between the context of a trial or an investigation (Sommers & Douglass, 2007).
4) Previous research has not examined the effects of evidence strength format on any guilt judgments for eyewitness identification (ID). Eyewitness ID evidence is one of the most persuasive types of evidence (e.g., Wells et al.,

2006), and thus it is critical to examine police accuracy in evaluating eyewitness ID evidence when judging a suspect's guilt.

5) There is mixed research on whether RMPs or LRs are superior formats for communicating evidence strength.

Chapter 3 contains Paper #2, which is a manuscript based on an experiment. Specifically, I recruited 209 U.S. police officers to complete the study on Qualtrics using CloudResearch, which is a company that can recruit from hard-to-reach populations. Police officers read a case scenario that indicated there is only 1 chance in 2 (fifty-fifty chance) that the suspect is guilty. Next, they read new evidence emerged and they read about one of nine pieces of evidence according to a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-subjects factorial design. The evidence strength format manipulated whether participants read about evidence in an LR, RMP, or neutral format. The evidence type factor manipulated whether participants read about DNA, fingerprint, or eyewitness ID evidence. Next, they answered two primary guilt measures: A 17-point log-scale measure and a percent likelihood scale. The log-scale served as the primary dependent variable based on previous research by Thompson and Newman (2015), whereas the percent likelihood scale served as an exploratory measure.

This study aimed to answer three primary research questions: 1) How does police accuracy in judging a suspect's guilt differ by evidence type; 2) How does police accuracy in judging a suspect's guilt differ by presentation format of objective evidence strength? 3) Do police judgments of suspect guilt vary by measurement type? I defined accuracy as the difference between participants' responses and the correct response.

Based on previous research, there were three main hypotheses.

Hypothesis 1a: Participants in the *DNA evidence* condition will be more accurate in their log-scale guilt judgments than those in the *fingerprint evidence* and *eyewitness evidence* conditions.

Hypothesis 1b: Participants in the *fingerprint evidence* condition will be more accurate in their log-scale guilt judgments than those in the *eyewitness evidence* conditions.

Hypothesis 2: Participants in the *RMP format* condition will be more accurate in their log scale guilt judgments than those in the *LR format* condition.

To investigate these hypotheses, I first created an accuracy measure by subtracting participants' log-scale responses from the correct response. Next, I conducted a factorial ANOVA on the accuracy measure using R. For exploratory analyses, I also conducted a factorial ANOVA on the log-scale (before converting to an accuracy measure) and the percent scale. The suggested publication outlet for Paper #2 is *Law and Human Behavior*. If publishing at *Law and Human Behavior* is unsuccessful, another suggested publication outlet is *Law, Probability and Risk*. I quantify my percentage effort for this article as 95% through conceptualizing the idea, developing a proposal to pursue this idea, developing the study design, applying for funding, recruiting participants, conducting the research and analyses, writing the manuscript, and interpreting the results. The remaining 5% effort is attributed to my co-author and advisor, Yueran Yang, for providing feedback on the design, materials, and the written manuscript.

Paper #3: Decreasing Biased Guilt Judgments During an Investigation with Social

Norms

In addition to understanding how police evaluate individual pieces of evidence when they judge suspect guilt, it is also important to understand how multiple pieces of evidence could influence police judgments of suspect guilt. Thus, another factor that could influence police guilt judgments is the order they encounter evidence. Based on previous literature, there are a few competing frameworks that can predict how evidence order affects police evaluations of evidence and judgments of suspect guilt.

I originally proposed competing hypotheses under the Bayesian Cognitive Model and confirmation bias. The Bayesian Cognitive Model suggests evidence order would not affect police evidence evaluations and judgments of suspect guilt (Druckman & McGrath, 2019; Edwards, 1962). The Bayesian Cognitive Model typically assumes people are motivated by accuracy to arrive at a correct conclusion (Druckman & McGrath, 2019) and that evaluations of new information are independent of prior beliefs or prior judgments (Blair & Rossmo, 2010; Druckman & McGrath, 2019). In other words, the Bayesian Cognitive Model assumes rationality.

Confirmation bias, alternatively, contradicts the Bayesian Cognitive Model of rationality and therefore could provide an alternative prediction as to how evidence order affects police judgments. Confirmation bias suggests that the first evidence police uncover during an investigation could bias evaluations of later evidence in the direction of the first piece of evidence (Nickerson, 1998). For example, if police investigators first uncover an eyewitness who identifies the suspect but later uncover inconclusive evidence, then they might still evaluate the second piece of evidence as incriminating due to confirmation bias and therefore render judgments of a suspect that are biased toward

guilt.

Although not originally predicted, data from the studies in Paper #3 suggested a recency bias rather than confirmation bias. A recency bias occurs when the last piece of evidence police uncover is more influential on guilt judgments than the first piece of evidence (Dahl et al., 2009; Carlson & Russo, 2001). For example, police might have higher guilt judgments of suspect guilt than if police discover incriminating evidence first and inconclusive evidence second than if they discovered inconclusive evidence first and incriminating evidence second. Previous research finds evidence of biased guilt judgments when mock jurors evaluate multiple pieces of evidence, but there are mixed results as to whether there are confirmation bias effects (e.g., Charman et al., 2017) or recency bias effects (Dahl et al., 2009).

One potential method to improve upon these biases is through social influence. The most prominent and instrumental form of social influence is conforming to relevant group norms (Hogg, 2010). Police investigators might look to their colleagues to gain information about group norms and to think consistently with their peers. Ask and colleagues (2011) found that social norms prioritizing a thorough investigation (vs. efficient) led to increased consideration of later evidence. That is, investigators in the thoroughness norms condition judged a suspect's guilt in the direction of the evidence discovered later in the case. Ask and colleagues (2011) suggested these results indicated police processed the later evidence more in the thoroughness condition than in the efficient condition, although their results might also suggest a recency effect. However, Ask and colleagues (2011) only manipulated incriminating and exculpatory evidence. They did not manipulate evidence order to determine whether thoroughness norms led

participants to be subject to a recency bias. Because biased effects are more likely to occur when there is ambiguous evidence, it is still unknown whether social norms could minimize bias. Police investigators adapting norms of thoroughness (vs. efficiency) should be less influenced by the order of evidence.

Thus, this paper attempts to address the following gaps:

1) Research examining the effects of evidence order on guilt judgments did not measure initial guilt beliefs (Charman et al., 2016; Dahl et al., 2009) or did not vary evidence order and compare initial guilt judgments to final guilt judgments (Charman et al., 2017).
2) Research examining the effects of evidence order on guilt judgments did not examine ambiguous evidence (Dahl et al., 2009; Price & Dahl, 2013), which is most prone to bias.
3) Research examining the effects of evidence order on guilt judgments was in the context of a criminal trial, rather than an investigation (Charman et al., 2016).
4) Research examining the effects of evidence order on guilt judgments examined potentially exonerating and incriminating evidence (Charman et al., 2016; Dahl et al., 2009), which could account for the discrepancy in that some researchers did not find confirmation bias (Charman et al., 2016; Dahl et al., 2009, Price & Dahl, 2013) whereas others found confirmation bias effects (Charman et al., 2017).
5) There is no research examining methods to minimize order effects (Charman et al., 2016; Dahl et al., 2009, Price & Dahl, 2013).

6) Social norms of thoroughness improved investigators' processing of two pieces of witness evidence (Ask et al., 2011). However, the authors did not vary the order of evidence.

Chapter 4 contains Paper #3, which is a manuscript based on two experiments. For the first experiment, I recruited 75 U.S. police officers by contacting police departments from my network, registering the study with the Professional Research Pool for Criminal Justice Science (n.d.), contacting police chiefs and personnel from a random sample of 200 departments from a list of 15,810 U.S. law enforcement agencies (LEAR, 2017), and posting the study on Police1, which is a news website with a police audience (police1.com). Due to time constraints, recruitment ended on December 31st. Due to the small sample size, I recruited a layperson sample (N = 636) using Amazon Mechanical Turk (MTurk) for the second experiment.

In both experiments, participants read a crime scenario describing a murder case. Next, I assigned participants to a 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (social norms: efficiency vs. thoroughness) x 2 (type: DNA incriminating vs. eyewitness incriminating) between-subjects factorial design. The order factor manipulated whether participants read about strongly incriminating evidence before or after ambiguous evidence. The type factor manipulated whether participants read about incriminating DNA evidence and ambiguous eyewitness ID evidence or incriminating eyewitness ID evidence and ambiguous DNA evidence. I included the evidence type factor to account for a potential confound between evidence type and evidence strength. The norms factor manipulated whether participants read their peers endorsed norms of investigative efficiency or thoroughness. Participants rated the

extent to which each piece of evidence was incriminating their judgments of suspect guilt after each piece of evidence. Participants then answered their perceptions of suspect and victim race (the crime scenario did not mention these attributes). The perceptions of suspect and victim race questions are reported in the Appendix at the end of the dissertation, as they are not included in the manuscript for Paper #3. Finally, participants completed a demographics questionnaire.

This study aimed to answer three primary research questions: 1) How does evidence strength format affect participants' accuracy in judging suspect guilt? 2) How does evidence type affect participants' accuracy in judging suspect guilt? 3) To what extent do participants' guilt judgments differ by guilt measure type? Based on the Bayesian cognitive framework and the confirmation bias, I originally predicted three competing hypotheses (see Figure 1).

Figure 1

Competing Hypotheses

Competing Hypotheses	**Bayesian Cognitive Model**	**Confirmation Bias**
H1	Mean final guilt judgments will be the same between order conditions	Final guilt judgments will be higher in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*)
H2	Incriminating evidence first condition, initial guilt judgments will not differ from mean final guilt judgments	Mean initial guilt judgments will be lower than mean final guilt judgments in the *incriminating first-ambiguous second* condition
H3	Evaluations of ambiguous evidence will be the same between order conditions	Evaluations of ambiguous evidence will be guiltier in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*)

Additionally, I predicted three hypotheses based on social norms theory.

Hypothesis 4: In the *incriminating evidence first-ambiguous second* condition, mean initial guilt judgments will be lower than mean final guilt judgments, but only among participants in the *efficiency norm* condition.

Hypothesis 5: Mean final guilt judgments will be higher in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*), but only among participants in the *efficiency norm condition* (vs. *thoroughness norm condition*).

Hypothesis 6: Social norms will moderate the effects of evidence order on

evaluations of ambiguous evidence, such that mean evaluations of ambiguous evidence will be guiltier in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*), but only among participants in the *efficiency norm condition*.

To investigate Hypotheses 1 and 4, I analyzed the effects of evidence order and social norms on final guilt judgments. For the first experiment, I conducted a between-groups factorial ANOVA in R. In the second experiment, I conducted a between-groups factorial ANCOVA in R to control for agreement to the norms statements because they were significantly related to the dependent measure among Amazon Mechanical Turk participants.

To investigate Hypotheses 2 and 5, I conducted a linear mixed-effects model in R to examine the effects of evidence order and social norms on the difference between initial and final guilt judgments. For the first experiment, fixed effects included the norms manipulation, evidence order manipulation, guilt judgment time (i.e., initial vs. final), and their interactions. Random effects included the participants' ID. For the second experiment, fixed effects included the norms manipulation, evidence order manipulation, guilt judgment type (i.e., initial vs. final), their interactions, evidence type, and agreement to the norms statements. I added the evidence type factor and participants' agreement with the norms statements as a fixed effect because these variables were significantly related to the dependent measure among Amazon Mechanical Turk participants (Study 2). Random effects included the participants' ID.

To investigate Hypotheses 3 and 6, I analyzed the effects of evidence order and social norms on evaluations of ambiguous evidence. For the first experiment, I conducted a between-groups factorial ANOVA in R. In the second experiment, I conducted a

between-groups factorial ANCOVA in R to control for agreement to the norms statements because they were significantly related to the dependent measure.

As mentioned previously, participants responded to their perceptions of the suspect's and victim's race. I did not include data from this question in the manuscript for Paper #3 due to this data being outside the scope of the paper, but the descriptive statistics for participants' responses are summarized as follows. First, most police participants in the first study that recruited police participants (N = 54, 87.1%) were "not sure" of the suspect's race. Some police participants reported the suspect was Black (N = 3, 4.8%), American Indian/Alaska Native (N = 2, 3.2%), or White (N = 2, 3.2%). Similarly, most police participants were "not sure" of the victim's race (N = 51, 82.3%), although reported the victim was White (N = 7, 11.3%), Black (N = 1, 1.61%), Asian (N = 1, 1.61%), or American Indian/Alaska Native (N = 1, 1.61%).

Second, most layperson participants in the second experiment (N = 340, 64.8%) were "not sure" of the suspect's race. Among layperson participants who reported their perceptions of suspect race, participants reported the suspect was White (N = 157, 29.9%), Black (N = 20, 3.8%), Asian (N = 6, 1.1%), American Indian/Alaska Native (N = 1, 0.2%), or "other" (N = 1, 0.2%). There was a similar trend for perceptions of victim race, as most layperson participants (N = 304, 57.9%) reported they were "not sure" of the victim's race. Among layperson participants who reported their perceptions of victim race, participants reported the suspect was White (N = 211, 40.2%), Black (N = 6, 1.14%), Asian (N = 3, 0.57%), or "other" (N = 1, 0.2%). Overall, most participants from both experiments correctly recalled there was no mention of suspect or victim race.

The data from these experiments suggested a recency bias, therefore an

application of recency bias to evidence order effects is included in the manuscript for Paper #3. The suggested publication outlet for Paper #3 is *Law and Human Behavior*. If publishing at *Law and Human Behavior* is unsuccessful, another suggested publication outlet is *Journal of Applied Social Psychology*. I quantify my percentage effort for this article as 95% through conceptualizing the idea, developing a proposal to pursue this idea, developing the study design, applying for funding, recruiting participants, conducting the research and analyses, writing the manuscript, and interpreting the results. The remaining 5% effort is attributed to my co-author and advisor, Yueran Yang, for providing feedback on the design, materials, and the written manuscript.

Dissertation Overview

The organization of this dissertation is as follows. Chapter 2 is comprised of Paper #1, Chapter 3 is comprised of Paper #2, and Chapter 4 is comprised of Paper #3. Each of these papers is a self-contained manuscript, therefore Chapters 2-4 include a title page, abstract, references, and appendices.

Chapter 5 discusses the major findings, contributions, implications, and limitations of the dissertation holistically by integrating each article into one cohesive narrative. The references section at the end of the dissertation only pertains to Chapter 1 and Chapter 5, as Chapters 2-4 have self-contained reference sections.

Chapter 2: Paper #1

Evaluating Evidence in a Criminal Investigation: Toward a Unified Definition of Evidence Strength

Jean J. Cabell[1] and Yueran Yang[2]

[1]Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno

[2]Psychology Department and Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno.

Author Note

Jean J. Cabell, https://orcid.org/0000-0002-2362-0419

Yueran Yang, https://orcid.org/0000-0001-9261-5608

We have no known conflicts of interest to disclose.

This work was supported by the Russell J. and Dorothy S. Bilinski Fellowship Fund, a program of the Bilinski Educational Foundation.

Correspondence concerning this article should be addressed to Jean J. Cabell, Interdisciplinary Social Psychology Ph.D. Program, Mailstop 1300, University of Nevada, Reno, 1664 N. Virginia St., Reno, NV 89557. Email: jcabell@nevada.unr.edu

Abstract

There are inconsistent definitions of evidence strength within psychological literature. Varying definitions of evidence strength could lead to inconsistency or inaccuracy when police use evidence to judge a suspect's guilt, as well as methodological problems (e.g., measurement error, ungeneralizable conclusions). There is also a need for greater nuance between how people actually evaluate evidence (subjective evidence strength) and how people should evaluate evidence (objective evidence strength) to assess the extent to which police are accurate when evaluating evidence to judge suspect guilt. We propose a definition of evidence strength based on competing hypotheses as a model to clarify evidence strength. We also propose using likelihood ratios and logarithm likelihood ratios to quantify objective evidence strength so researchers can define objective evidence strength. Finally, we discuss several areas for future research that will aid in consistently operationalizing evidence strength and provide greater nuance between subjective and objective evidence strength.

Evaluating Evidence in a Criminal Investigation: Toward a Unified Definition of Evidence Strength

Evidence is a critical component of any police investigation because police rely on evidence to infer suspect guilt to make subsequent decisions. For example, police might use evidence to judge a suspect's guilt before deciding whether they need to collect more evidence, make an arrest, refer a case to the prosecution, or pursue a different suspect. However, there is a lack of a consensus to define and measure evidence strength within psychological literature. This lack of a uniform operational definition could lead to measurement error and research that lacks generalizable conclusions. More importantly, varying operationalizations of evidence strength could lead to inconsistency or inaccuracy when police use evidence to infer suspect guilt if their decisions are based on research that does not have a unified operationalization of evidence strength.

The purpose of this paper is to clarify the meaning of evidence strength. Specifically, we use competing hypotheses to define how a piece of evidence conveys evidence strength. We also propose greater nuance between how police *subjectively* evaluate evidence strength compared to an *objective* standard of evidence strength. Such a distinction between subjective and objective evidence strength is necessary to assess accuracy in how police evaluate evidence strength, as researchers must determine to what extent subjective evaluations of evidence strength align with objective evidence strength. We conclude by discussing several implications for future research. Although our conclusions could be expanded to other areas of the criminal justice system (e.g., judges, juries, plea bargaining), we will be limiting our application to the context of police investigations for the sake of scope and brevity.

Inconsistencies in Operationalizing Evidence Strength

Previous psychological literature varies in operationalizing evidence strength, suggesting there is not a consensus on how researchers should define and measure evidence strength. This lack of a consensus is problematic in that police could be inconsistent or inaccurate when using evidence to infer suspect guilt. Although the reviewed literature is not exhaustive of all psychological research on evidence strength, this section serves to provide a basis for our argument that there are inconsistent operationalizations of evidence strength.

One definition of evidence strength is *the extent to which a piece of evidence is incriminating*. For example, Lidén and colleagues (2019a, 2019b) measured evidence strength by asking participants to rate the extent to which a piece of evidence suggested a defendant was guilty from 1 (*very weakly*) to 7 (*very strongly*). However, operationalizing evidence strength as *the extent to which a piece of evidence is incriminating* can be problematic because it does not consider competing hypotheses. This definition implies that police are only concerned with one hypothesis: *The person is guilty* (H1). However, any hypothesis must be accompanied by at least one competing hypothesis (i.e., an alternative hypothesis). The competing hypothesis for H1 could be either *the person is innocent* (H2a) or *the person is not guilty* (H2b) in this example. Although these two alternative hypotheses might sound semantically similar, they convey different information. For example, if a researcher used a Likert-type scale from 1 to 7 and asked participants to rate their agreement with the statement *the piece of evidence is incriminating*, it is unknown whether a “1” response suggests the participant believes the suspect is innocent or whether the participant believes the suspect is equally guilty or

innocent. In other words, defining evidence strength only in terms of the extent to which the evidence is incriminating neglects considering the valence of the evidence. Evidence in this example would only be "strong" when it supports propositions of guilt, but all exculpatory evidence and ambiguous evidence would be considered weak, regardless of valence.

Again, failing to explicitly state the alternative hypothesis creates difficulty in evaluating the extent to which the evidence provides support for one hypothesis versus another. Therefore, the legal system and scholars ought to define evidence strength as *the extent to which the evidence supports one hypothesis versus another hypothesis.* However, which set of competing hypotheses should researchers and the criminal justice system use to define evidence strength? The set of *the person is guilty* (H1) and *the person is innocent* and *the person is innocent* (H2a) provides a positively stated alternative hypothesis, whereas *the person is guilty* (H1) and *the person is not guilty* (H2b) provides a negatively stated alternative hypothesis.

The legal system tends to examine the competing hypotheses of *the person is guilty* (H1) vs *the person is not guilty* (H2b), as courts only allow people to be legally "guilty" or "not guilty[1]." Suspects and defendants cannot plead or be found "innocent" to alleged crimes. However, as indicated by the above example, this set of competing hypotheses is problematic during an investigation because all ambiguous and exculpatory evidence would be "weak." The negatively stated alternative hypothesis *the person is not guilty* (H2b) fails to capture the valence of the evidence and cannot distinguish between

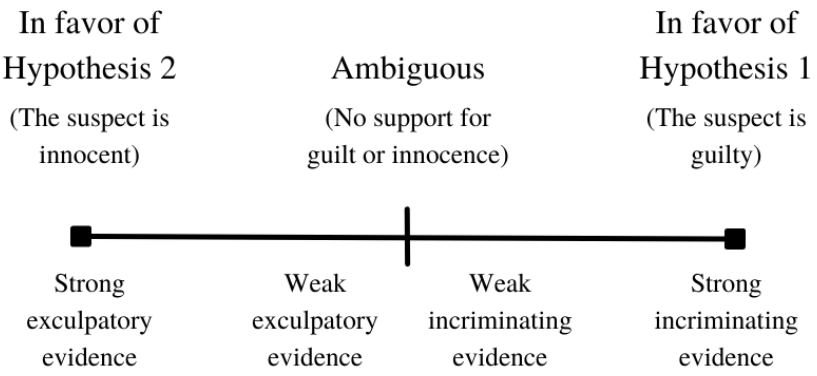
[1] Defendants can also plead "no contest" such that they do not formally admit guilt, but they agree to the punishments for the alleged crime.

ambiguous and exculpatory evidence.

Thus, we consider it more appropriate to define evidence strength as *the extent to which the evidence supports the hypothesis "the person is guilty"* (H1) *versus the alternative hypothesis "the person is innocent"* (H2; i.e., *the extent to which the evidence is incriminating versus exculpatory*). As shown in Figure 1, this definition operationalizes evidence strength as a continuum between the two positively stated competing hypotheses and thus can embrace the entire range of evidence strength. Note that this definition can account for both the valence and magnitude of evidence strength. The valence reflects which hypothesis the evidence is supporting, in other words, whether the evidence is incriminating (supporting *the suspect is guilty* [H1]) or exculpatory (supporting *the suspect is innocent* [H2]). The magnitude reflects how strong the evidence is in supporting the corresponding hypothesis. In other words, evidence can be strong in supporting either H1 (strongly incriminating) or supporting H2 (strongly exculpatory). Evidence is weak when it does not provide support to either hypothesis.

Figure 1

Visual Depiction of Evidence Strength Definition



Subjective versus Objective Evidence Strength

In addition to clarifying a definition of evidence strength, there is also a need to distinguish between *perceived* evidence strength and *actual* evidence strength. Such a distinction is necessary to assess police's accuracy in evaluating evidence strength. To illustrate the importance of this distinction, consider the following example. People generally perceive eyewitness ID evidence as strong, such that they tend to believe suspects or defendants are guilty when there is incriminating eyewitness ID evidence (National Institute of Justice, 1999; Wells et al., 2006; Wells & Olsen, 2011). However, police do not always follow "pristine condition" guidelines that preserve the integrity of eyewitness ID evidence (see Wells et al., 2020 for a comprehensive review of the pristine conditions to collect eyewitness evidence) and eyewitnesses can be inaccurate. Eyewitness ID evidence collected without pristine conditions might be subjectively strong in that police *perceive* the evidence as strongly incriminating, but the evidence is

verifiably weak or ambiguous because it cannot offer reliable information regarding suspect guilt. Therefore, there must be a distinction between *subjective evidence strength* (perceived evidence strength based on individual opinions) and *objective evidence strength* (actual evidence strength based on a verifiably observed truth) because they refer to two separate constructs.

Distinguishing between objective and subjective evidence strength in the literature is useful for research, policy, and practice because it offers nuance between descriptive and normative definitions. Cognitive research often separates descriptive and normative standards to distinguish between how people *actually* think and make decisions (descriptive) versus how people *ought* to think and make decisions (normative; Over, 2004). Thus, subjective evidence strength refers to a descriptive definition of how people actually perceive evidence strength, whereas objective evidence strength refers to a normative definition of how people ought to perceive evidence strength. Researchers can then compare these two separate constructs to assess the extent to which police are accurate in evaluating evidence strength.

Quantifying Objective Evidence Strength

As stated in the first section, we propose to clarify evidence strength by explicitly stating two competing, positively stated hypotheses. We also distinguish between subjective and objective evidence strength. As a normative standard, how could the legal system and scholars quantify objective evidence strength? Below we discuss some common methods to measure objective evidence strength, including likelihood ratios, random match probabilities, and logarithm likelihood ratios.

Likelihood Ratios

Likelihood ratios are one method to quantify evidence strength. Likelihood ratio (LR), also called the Bayes factor (Bolstad & Currant, 2016) or diagnostic ratio (Wells & Lindsay, 1980), is the ratio between two probabilities. Often in psychology and law research, this probability is that the evidence will occur under H1 (the suspect is guilty) and the probability that the evidence will occur under H2 (the suspect is innocent). Therefore, LRs often measure evidence strength by comparing a ratio between the extent to which the evidence supports the hypothesis that the suspect is guilty over the hypothesis that the suspect is innocent.

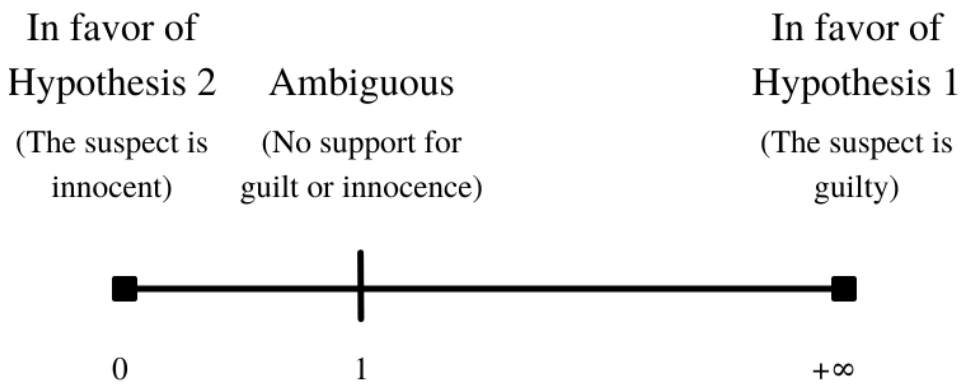
As shown in Figure 2, an LR of 1 indicates the evidence does not provide any support to either hypothesis. Above the value of 1, the higher the LR, the stronger the eyewitness ID evidence is at supporting the hypothesis that the suspect is guilty of committing the crime. Below the value of 1, the lower the LR, the stronger the eyewitness ID evidence is at supporting the hypothesis that the suspect is innocent of committing the crime.

LRs commonly quantify evidence strength. For example, researchers (e.g., Steblay et al., 2011; Wells & Lindsay, 1980) have used LRs to represent how much more likely it is that an eyewitness identified a suspect from a lineup given the suspect is guilty (Hypothesis 1) compared to innocent (Hypothesis 2). The higher the LR, the stronger the eyewitness ID evidence is at supporting the hypothesis that the suspect is guilty of committing the crime (see Figure 1). Researchers have also used LRs to convey the strength of forensic evidence (Meuwly et al., 2017; Thompson et al., 2018) and confession evidence (Horgan et al., 2012; Russano et al., 2005) in addition to eyewitness

ID evidence.

Figure 2

Quantifying Objective Evidence Strength using Likelihood Ratios



Note. The figure above is a visual depiction of how likelihood ratios can be used to quantify evidence strength (adapted from de Keijeser & Elffers, 2012; Martire et al., 2014). This figure is not to scale.

The *lineup-as-experiment* analogy (Wells & Luus, 1990) can clarify why LRs are useful for quantifying evidence strength. This analogy posits that the eyewitness lineup task is similar to a psychology experiment because both involve a procedure to test a hypothesis. The outcomes of both the eyewitness task and a psychology experiment are "probabilistic in their discovery of truth" given that their outcomes will only ever be a statistical truth rather than the ground truth, even under the most pristine conditions (Wells & Luus, 1990, pp. 107). For example, false identifications can occur even if an eyewitness chooses the suspect from a lineup conducted under pristine conditions. Likewise, an experiment could result in a Type I or Type II error even when methodologically flawless. The collection of other police evidence is similar to the eyewitness identification task, as well as to a psychology experiment. Even when police collect evidence under the best of circumstances, there is still a possibility for error. Thus,

LRs can express probabilistic uncertainty and convey the extent to which a piece of evidence supports the competing hypotheses that police are attempting to test during a criminal investigation.

Random Match Probabilities

LRs might be difficult to interpret, as people tend to understand probabilities better when presented as a frequency than a ratio (Gigerenzer & Hoffrage, 1995). Random match probabilities (RMPs) were developed to facilitate people's understanding of LRs because RMPs use a frequency format to convey the same information as LRs (Thompson et al., 2018; Thompson & Newman, 2015). For example, fingerprint evidence presented in an RMP format would be presented as *one person in 10 would have a fingerprint consistent with the fingerprint sample from the crime scene.* Conversely, an LR would be conveyed as *it is 10 times more likely the evidence occurred given the fingerprint originated from the suspect versus a randomly chosen person.*

People can sometimes interpret RMPs more easily than LRs, as mock jurors more accurately interpreted shoeprint evidence strength when the evidence strength was presented as an RMP than as an LR (Thompson & Newman, 2015). However, people do not always interpret RMPs more accurately than LRs, as sometimes they interpret RMPs similarly or even less accurately than LRs. People tend to interpret RMPs similarly to LRs for DNA evidence, for instance (Thompson et al., 2018; Thompson & Newman, 2015). Sometimes people even interpreted RMPs as conveying the opposite evidence strength than what they were intended to convey (Thompson et al., 2018).

RMPs can also be misleading in some situations. For example, LRs are preferred when two forensic samples have the same source but the features necessary for

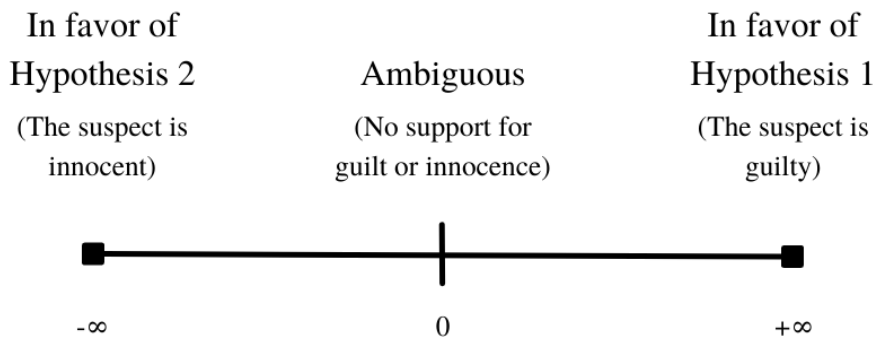
comparison are not guaranteed to be observed (see Thompson et al., 2018). Imagine two DNA samples have the same source, but one DNA sample only contains a partial profile whereas the other sample contains a full profile. In this case, the features necessary for comparison are not guaranteed to be observed in both samples. LRs are preferred over RMPs in this instance because the RMP no longer conveys the same information as the LR (i.e., $p[E|H] < 1.0$). Otherwise, LRs and RMPs do convey the same information (i.e., $p[E|H] = 1.0$).

Logarithm Likelihood Ratios

Given that LRs can be difficult to interpret and RMPs are not always superior, another format to ease the interpretation of objective evidence strength is the logarithm of the LR (log LR). A logarithm in base 10 represents the power of 10 that a number represents. In other words, a logarithm in base 10 refers to how many digits that number contains. The logarithm of an LR of 10 (10^1) is 1, 100 (10^2) is 2, 1000 (10^3) is 3, and so forth (see Figure 3).

Figure 3

Quantifying Objective Evidence Strength using Logarithm Likelihood Ratios



Note. The figure above depicts the range for quantifying evidence strength using a logarithm of a

likelihood ratio.

This format of quantifying objective evidence strength has several benefits. First, the neutral point of a log LR is 0, as compared to the neutral point of an LR, which is 1. The neutral point of a log LR being 0 is beneficial because it is easier to identify which hypothesis the evidence is supporting by whether the number is negative or positive. There are no negative values with a traditional LRs and thus it is not intuitive which hypothesis the evidence supports. For example, an LR of 10 (10^1) has the same magnitude as an LR of 0.1 (10^{-1}), but this equivalence is not necessarily intuitive. Rather, it is easier to understand that a log LR of -1 and 1 have same the magnitude but are opposite valences. Thus, log LRs provide a natural way to represent objective evidence strength that conveys the magnitude and valence of any given piece of evidence (Robertson et al., 2016).

Second, a log LR provides a more intuitive method to calculate the likelihood of suspect guilt when there are multiple pieces of evidence compared to a traditional LR. With traditional LRs it is necessary to use multiplication to quantify total evidence strength when there are multiple pieces of evidence (Robertson et al., 2016). For example, if the LR of fingerprint evidence is 100 and the LR of the eyewitness evidence is 10, then it is necessary to multiply 100 by 10 to calculate an LR of 1000. Log LRs require simpler, additive properties: $10^1 \times 10^2 = 10^3$, or $1 + 2 = 3$. Thus, using log LRs to calculate total evidence strength for multiple pieces of evidence simply requires addition and subtraction, which is much easier to calculate and understand when considering neutral evidence or evidence that supports the alternative hypothesis.

Measuring Objective and Subjective Evidence Strength

To compare subjective evaluations of evidence to objective evidence strength, there must be established methods to calculate objective evidence strength and to measure subjective evidence strength. This section first reviews existing calculations of objective evidence strength. This section follows by describing a need to establish objective evidence strength, accurately measure evidence strength, and conduct research comparing subjective evidence strength to objective evidence strength within the context of a police investigation.

Objective Evidence Strength Measurements by Evidence Type

Researchers have primarily used LRs and RMPs to statistically define objective evidence strength for three main classes of evidence: eyewitness, confession, and forensic evidence. This section will review the similarities and differences between objective evidence strength by evidence type.

Objective Eyewitness Evidence Strength

One of the most common methods for conveying the accuracy of eyewitness ID evidence is through the diagnosticity ratio. The diagnosticity ratio is a likelihood ratio (LR) between accurate and mistaken eyewitness IDs (Wells & Lindsay, 1980). This ratio represents how much more likely an eyewitness ID (or non-ID) was to have occurred given the truth of one hypothesis (e.g., that the suspect is guilty) relative to another hypothesis (e.g., that the suspect is innocent). The higher the diagnosticity ratio, the stronger the eyewitness ID evidence is in supporting the hypothesis that the suspect is the culprit (Steblay et al., 2011).

Steblay and colleagues (2011) conducted a meta-analysis of 72 studies that

calculated diagnosticity ratios and found that eyewitness ID evidence is strongest when an eyewitness makes an ID from a sequential lineup (vs. a simultaneous lineup). Therefore, diagnosticity ratios suggest that eyewitness ID evidence is stronger incriminating evidence when police use a sequential lineup versus a simultaneous lineup (though see Wixted & Mickes, 2015 for a critical review of diagnosticity ratios).

Objective Confession Evidence Strength

Diagnosticity ratios (i.e., LRs) have also been applied to confession evidence (e.g., Horgan et al., 2012; Russano et al., 2005). When applied to confession evidence, diagnosticity ratios typically represent the ratio of true to false confessions as related to various police interrogation techniques (Meissner et al., 2010). In other words, diagnosticity ratios for confession evidence convey the likelihood a suspect is guilty versus innocent given a confession, depending upon which interrogation techniques the police used. Diagnosticity ratios were lower among police who used the techniques of minimization, maximization, and offers of leniency during an interrogation (Russano et al., 2005), whereas diagnosticity ratios were higher among police who presented true evidence and emphasized the morality of confession increased the diagnosticity ratio (Horgan et al., 2010). Therefore, confessions are weaker, incriminating forms of evidence when police use minimization, maximization, or offers of leniency as compared to police who present true evidence or persuade suspects that it is moral to confess.

Objective Forensic Evidence Strength

The psychological literature on LRs for eyewitness and confession evidence primarily focuses on factors (e.g., confidence, lineup type) that are related to the reliability and accuracy of eyewitness ID evidence. However, much of the psychological

literature on statistical classifications of forensic evidence focuses on how people interpret different formats of LRs, rather than on factors that influence the reliability and accuracy of forensic evidence.

LRs are, by default, expressed numerically. For example, weak incriminating fingerprint evidence would be expressed as *it is 10 times more likely that the two fingerprints originated from the same person than from different people* (Martire et al., 2013; Martire et al., 2014). Some organizations are proponents of using a *verbal* equivalent of LRs under the assumption that verbal expressions of LRs are easier to understand than numerical expressions of LRs (e.g., Association of Forensic Science Providers, 2009). For example, a verbal equivalent of an LR of 10 would be conveyed as *there is weak to limited support that the two fingerprints originated from the same person than from different people* (Martire et al., 2013; Martire et al., 2014). However, most research finds numerical expressions of LRs are still superior when compared to verbal expressions or verbal expressions should be expressed in conjunction with numerical expressions to convey appropriate evidence strength (Marquis et al., 2016; Martire et al., 2014; Thompson & Newman, 2015).

Establishing Objective Evidence Strength

The first step in identifying how people ought to be evaluating evidence is to establish the objective evidence strength of any given type of evidence. Although established methods exist for calculating LRs for DNA (e.g., Lohmueller & Rudin, 2013), fingerprint (e.g., Ramos et al., 2017), eyewitness ID (e.g., Wells & Lindsay, 1980), and confession evidence (e.g., Horgan et al., 2012; Russano et al., 2005), there are several types of evidence that lack a method to calculate objective evidence strength. For

example, alibi evidence could benefit from research that establishes factors that contribute to its diagnosticity, similar to how researchers have established factors that affect LRs for eyewitness ID (e.g., Wells & Lindsay, 1980) and confession evidence (e.g., Horgan et al., 2012; Russano et al., 2005).

A method to calculate LR alone is not sufficient because the objective evidence strength of forensic evidence can be miscalculated due to human error (Jang, 2021; National Research Council, 2009). Further, forensic science can be prone to testing errors, as such testing errors were among the most important factors that contributed to wrongful convictions (Saks & Koehler, 2005). This human error can occur during the collection, assessment, and interpretation of forensic evidence (e.g., Jang, 2021; Kaplan et al., 2020), rendering a need for identifying what factors lead to pristine conditions to collect, assess, and interpret forensic evidence.

Consider fingerprint evidence as an example. Methods for calculating the LRs for fingerprint evidence exist but determining the extent to which two fingerprints are similar is still a subjective task. Because fingerprint evidence still involves subjective methodologies, one option could be to use "evidence lineups" with forensic evaluators who are comparing fingerprint evidence (Kassin et al., 2013; Kukucka et al., 2020; Quigley-McBride & Wells, 2018). These evidence lineups would involve embedding the suspect's sample among known-innocent filler samples.

Kukucka and colleagues (2020) recruited forensic examiners and found evidence lineups affected forensic examiners' decision-making similarly to how eyewitness lineups affect eyewitness decision-making, suggesting forensic science could benefit from evidence lineups to improve the diagnosticity of forensic evidence. Future research

could focus on identifying the pristine conditions for any type of forensic evidence that involves subjective comparisons (e.g., shoeprint, hair analysis, handwriting) so that their measurements of objective evidence strength are more likely to be valid.

Measuring Subjective Evidence Strength

Researchers interested in comparing the extent to which subjective evidence strength aligns with objective evidence strength need accurate measures of subjective evidence strength. Likert-type scales are one method to measure subjective evidence strength. For example, Greenspan and Scurich (2016) measured participants' subjective evidence strength of eyewitness and alibi evidence on a 9-point Likert-type scale. Such a measure is useful for quantifying subjective evidence strength when the goal is to describe subjective evidence strength (Greenspan & Scurich, 2016). However, this measure could be problematic if the goal is to compare subjective evidence strength to objective evidence strength for three reasons. First, it can be difficult to compare results from different studies if researchers measure subjective evidence strength on Likert-type scales with different ranges (e.g., 9-point vs. 7-point). Second, these scales limit the range of answers to a range that cannot encompass the full range of LRs or log LRs. Third, Likert-type scales are not naturally comparable to LRs and log LRs, thereby inducing difficulty in using Likert-type scales comparing subjective to objective evidence strength to assess accuracy.

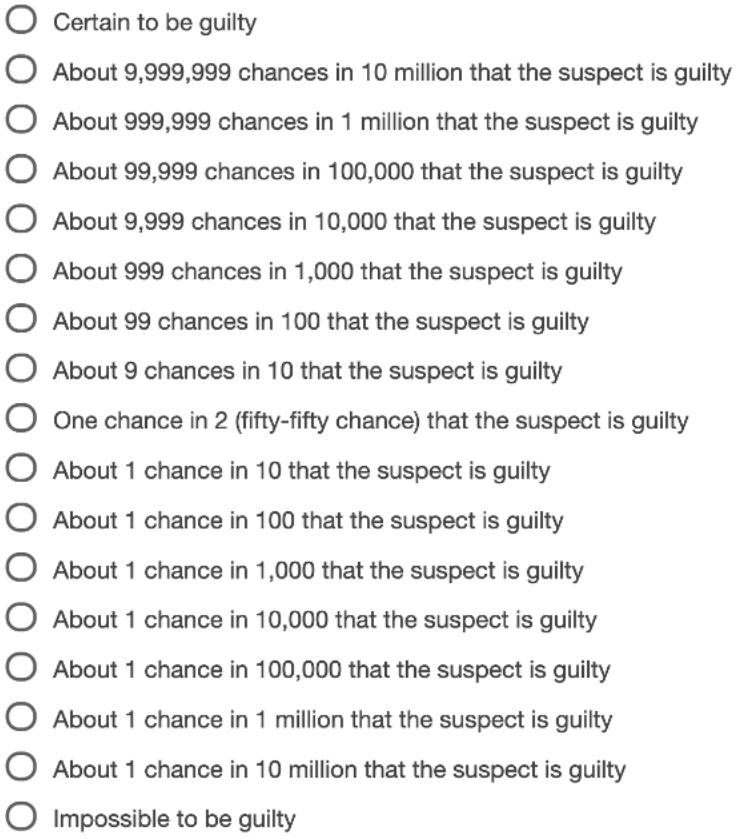
There are examples of using a larger range of response options to measure evidence strength to compare subjective to objective evidence strength. For example, Martire and colleagues (2014) used a statement of odds scale to measure subjective evidence strength. Participants who believed an accused defendant was more likely guilty

than not guilty entered a number greater than 1 in the following statement: *Based on the available evidence I believe that it is ____ times more likely that the accused is guilty than not guilty*. Using this method to measure subjective evidence strength theoretically can encapsulate any number and can be compared to LRs.

However, participants' subjective evidence strength was more accurate on a 17-point log likelihood scale than on a statement of odds (Thompson & Newman, 2015) suggesting a log-scale could be an appropriate method to measure subjective evidence strength. Thompson and Newman (2015) used a 17-point log-scale to measure subjective evidence strength by asking participants to fill in the following sentence with their perceived chances of suspect guilt: "Based on the available evidence, I believe the chances the suspect is guilty of committing murder is ___." Items on this scale range from "Certain to be guilty" to "Impossible to be guilty", with each interval being approximately equal on a scale of log odds (see Figure 4 for an example). A log-scale facilitates expressing high and low values of evidence strength. However, this scale hinders expressing mid-point values of subjective evidence strength due to its concentration on high and lower values of evidence strength. This scale also used a negatively stated alternative hypothesis, which is problematic for precisely measuring evidence strength, particularly because it was measuring objective evidence strength that was based on two positively stated competing hypotheses. Future research could examine what measures of subjective evidence strength are best when compared against objective evidence strength.

Figure 4

Log-scale Measure of Guilt Judgments



Note. The figure above depicts an example of a 17-point log-scale.

Comparing Subjective Evidence Strength to Objective Evidence Strength

Ideally, police's subjective evidence strength should align with objective evidence strength to minimize judgment errors. However, there are many examples of a misalignment between subjective and objective evidence strength. For example, police might discount exculpatory evidence—especially when they already have expectations of guilt. Indeed, officers rated inconsistent witness evidence as less reliable than consistent witness evidence when they had prior expectations of guilt (Ask & Granhag, 2007).

Another context where subjective evidence strength does not align with objective evidence strength is when there is weak evidence. Weak evidence can sometimes be subject to the *weak evidence effect* (e.g., Martire et al., 2014; Martire et al., 2013). The weak evidence effect refers to when participants' subjective evidence strength is in the opposite direction than what the objective evidence strength conveys. For example, participants interpreted weakly incriminating DNA evidence as exculpatory, despite the objective evidence strength of the evidence being incriminating (Martire et al., 2014).

Still, there are relatively few studies that attempt to identify whether people are accurate in their subjective evaluations of evidence strength, signifying a need to identify instances where the two do not align. Particularly, there is little research that identifies the extent to which police are accurate in their subjective evaluations of evidence strength. It is necessary to determine where there are errors in subjective evidence strength aligning with objective evidence strength because these errors could lead to the criminal justice system not accomplishing its primary goal: correctly convicting the guilty.

There are a few understudied areas that could benefit from identifying where these errors between subjective and objective evidence strength occur. First, the extent to which subjective and objective evidence strength aligns has differed by evidence type (e.g., Thompson & Newman, 2015), but there are still many types of evidence that have yet to be examined. For example, confession evidence is another persuasive type of evidence (Kassin et al., 2013) but there is a lack of research examining how accurately police interpret this evidence when presented in a statistical format. There is also little research that evaluates how accurately police or laypeople interpret other types of

forensic evidence (e.g., handwriting, hair analysis, bitemark). Some research has estimated fingerprint evidence is more powerful than eyewitness, which is more powerful than confession evidence (Blair & Rossmo, 2010), so it is possible that police might subjectively perceive some evidence types to be more incriminating even when the evidence has the same objective evidence strength.

Second, errors in evaluating evidence strength have differed by whether participants read the evidence in an LR format or an RMP format (Thompson & Newman, 2015), but there is still mixed evidence as to which format people understand most accurately. For example, Thompson and colleagues (2018) found some people misinterpreted RMPs in a pilot study but they also found people interpreted LRs similarly to RMPs in their third study. There is also no research to our that examines how police or laypeople interpret log LRs, especially compared to other statistical formats of objective evidence strength. Although log LRs are likely easier to understand (Robertson et al., 2016), this intuition has yet to be empirically examined. Identifying the best format by which to communicate objective evidence strength is important to understand how best to align subjective evidence strength with objective evidence strength during criminal investigations.

Conclusion

Many fields examine how people evaluate varying strengths of evidence and how these evaluations inform judgments of suspect guilt without a consistent operationalization of evidence strength. As previously reviewed, there are varying operationalizations of evidence strength. This lack of consensus could lead to measurement error and a lack of generalizable conclusions in research, as well as

inconsistency or inaccuracy when police use evidence to infer suspect guilt. Our main takeaway messages are as follows: (1) Researchers should aim for one overall definition of evidence strength, (2) researchers should distinguish between objective evidence strength and subjective evidence strength within their research. For researchers defining and examining objective evidence strength, log LRs are likely the most accurate and intuitive formats to communicate objective evidence strength.

This paper offers several implications, including a need: 1) To establish objective evidence strength with several types of evidence, 2) for an accurate measure of subjective evidence strength, 3) for more research on comparing subjective evidence strength to objective evidence strength among police. Our goal for this paper is to begin a broader discussion for a consistent definition of evidence strength, as well as to distinguish between objective and subjective evidence strength because they are different constructs that accomplish different research goals. Such a clarification will be useful for future research and practice to ultimately work toward the goal of police accurately evaluating evidence to judge suspect guilt during police investigations.

References

Ask, K., & Granhag, P. A. (2007). Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology*, *37*(3), 561–591. https://doi.org/10.1111/j.1559-1816.2007.00175.x

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*, 161–164. https://doi:10.1016/j.scijus.2009.07.004

Blair, J. P., & Rossmo, D. K. (2010). Evidence in context: Bayes' theorem and investigations. *Police Quarterly*, *13*(2), 123–135. https://doi.org/10.1177/1098611110365686

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics* (3rd ed.). John Wiley & Sons.

de Keijser, J., & Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law*, *18*(2), 191–207. https://doi.org/10.1080/10683161003736744

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Greenspan, R., & Scurich, N. (2016). The interdependence of perceived confession voluntariness and case evidence. *Law and Human Behavior*, *40*(6), 650–659. https://doi.org/10.1037/lhb0000200

Horgan, A. J., Russano, M. B., Meissner, C. A., & Evans, J. R. (2012). Minimization and maximization techniques: Assessing the perceived consequences of confessing

and confession diagnosticity. *Psychology, Crime & Law*, *18*(1), 65–78. https://doi.org/10.1080/1068316X.2011.561801

Jang, M. (2021). *Impacts of evidence on decision-making in police investigation* (Doctoral dissertation). Retrieved from Electronic Theses Online Service (EThOS). (Order No. uk.bl.ethos.837277)

Kaplan, J., Ling, S., & Cuellar, M. (2020). Public beliefs about the accuracy and importance of forensic evidence in the United States. *Science & Justice*, *60*(3), 263–272. https://doi.org/10.1016/j.scijus.2020.01.001

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Kukucka, J., Dror, I. E., Yu, M., Hall, L., & Morgan, R. M. (2020). The impact of evidence lineups on fingerprint expert decisions. *Applied Cognitive Psychology*, *34*(5), 1143–1153. https://doi.org/10.1002/acp.3703

Lidén, M., Gräns, M., & Juslin, P. (2019a). From devil's advocate to crime fighter: Confirmation bias and debiasing techniques in prosecutorial decision-making. *Psychology, Crime & Law*, *25*(5), 494–526. https://doi.org/10.1080/1068316X.2018.1538417

Lidén, M., Gräns, M., & Juslin, P. (2019b). 'Guilty, no doubt': Detention provoking confirmation bias in judges' guilt assessments and debiasing techniques. *Psychology, Crime & Law*, *25*(3), 219–247. https://doi.org/10.1080/1068316X.2018.1511790

Lohmueller, K. E., & Rudin, N. (2013). Calculating the Weight of Evidence in Low-

Template Forensic DNA Casework. *Journal of Forensic Sciences*, *58*(s1), S243–S249. https://doi.org/10.1111/1556-4029.12017

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F., & Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, *56*(5), 364–370. https://doi.org/10.1016/j.scijus.2016.05.009

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, *240*, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, *37*(3), 197–207. https://doi.org/10.1037/lhb0000027

Meissner, C. A., Hartwig, M., & Russano, M. B. (2010). The need for a positive psychological approach and collaborative effort for improving practice in the interrogation room. *Law and Human Behavior*, *34*(1), 43–45. https://doi.org/10.1007/s10979-009-9205-9

Meuwly, D., Ramos, D., & Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, *276*, 142–153. https://doi.org/10.1016/j.forsciint.2016.03.048

National Institute of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: Office of Justice Programs, U.S. Department of Justice

National Research Council (U.S.), National Research Council (U.S.), & National Research Council (U.S.) (Eds.). (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press.

Over, D. E. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3–18). Malden, MA: Blackwell Publishing.

Quigley-McBride, A., & Wells, G. L. (2018). Fillers can help control for contextual bias in forensic comparison tasks. *Law and Human Behavior*, *42*(4), 295–305. https://doi.org/10.1037/lhb0000295

Ramos, D., Haraksim, R., & Meuwly, D. (2017). Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data in Brief*, *10*, 75–92. https://doi.org/10.1016/j.dib.2016.11.008

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons.

Russano, M. B., Meissner, C. A., Narchet, F. M., & Kassin, S. M. (n.d.). *Investigating true and false confessions within a novel experimental paradigm*. *16*(6), 6.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895. https://doi.org/10.1126/science.1111565

Thompson, W. C., & Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, *39*(4), 332–349.

https://doi.org/10.1037/lhb0000134

Thompson, W. C., Grady, R. H., Lai, E., & Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, *17*(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., & Luus, C. A. E. (1990). Police lineups as experiments: Social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin*, *16*(1), 106–117. https://doi.org/10.1177/0146167290161008

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*(1), 277–295. https://doi.org/10.1146/annurev.psych.54.101601.145028

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*(1), 3–36. https://doi.org/10.1037/lhb0000359

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*(2), 45–75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*(2), 262–276.

https://doi.org/10.1037/a0035940

Chapter 3: Paper #2

Police Comprehension of Evidence: Statistical Formats and Evidence Type

Jean J. Cabell[1] and Yueran Yang[2]

[1]Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno

[2]Psychology Department and Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno.

Author Note

Jean J. Cabell, https://orcid.org/0000-0002-2362-0419

Yueran Yang, https://orcid.org/0000-0001-9261-5608

We have no known conflicts of interest to disclose.

This work was supported by the Russell J. and Dorothy S. Bilinski Fellowship Fund, a program of the Bilinski Educational Foundation.

Correspondence concerning this article should be addressed to Jean J. Cabell, Interdisciplinary Social Psychology Ph.D. Program, Mailstop 1300, University of Nevada, Reno, 1664 N. Virginia St., Reno, NV 89557. Email: jcabell@nevada.unr.edu

Abstract

Objective: Police must properly evaluate evidence to make consequential decisions, but erroneous evaluations of evidence could begin a process that leads to wrongful convictions through faulty guilt judgments. This study aimed to assess the extent to which police accuracy in judging suspect guilt varied by the format of the evidence strength and the type of evidence. **Hypotheses:** We predicted police would be most accurate in judging suspect guilt when we presented evidence strength in a random match probability (RMP) format compared to a likelihood ratio (LR) format. We also predicted police would be more accurate at evaluating DNA evidence (vs. fingerprint and eyewitness ID) and fingerprint evidence (vs. eyewitness ID). **Method**: We recruited police participants ($N = 209$) using CloudResearch and randomly assigned them to a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-subjects factorial design. **Results**: Surprisingly, police under-weighed evidence in the RMP format and were most accurate at judging suspect guilt when the evidence was in the LR format, but this effect only occurred for DNA evidence. Police sometimes over-weighed or under-weighed DNA evidence compared to fingerprint evidence, but this effect differed by evidence strength format. Overall, LRs produced the most accurate shifts in guilt judgments. **Conclusion**: Forensic experts and researchers should use LRs to convey evidence strength when possible.

Police Comprehension of Evidence: Statistical Formats and Evidence Type

Over 2,900 people in the U.S. have been wrongfully convicted of crimes they did not commit, collectively spending over 26,500 years erroneously imprisoned (The National Registry of Exonerees, 2022). One factor that might lead to wrongful convictions is whether police can accurately evaluate evidence during an investigation. Particularly, the extent to which police investigators can accurately interpret evidence determines whether they can accurately judge a suspect's guilt. If police investigators misinterpret the evidence, they might misidentify innocent people as suspects, which could ultimately lead to wrongful convictions (Scherr et al., 2020). Thus, evaluating evidence properly throughout a criminal investigation is a critical factor to avoid wrongfully incriminating an innocent person. Further, evaluating evidence properly is critical to correctly solve a case and incriminate a guilty suspect. This study focused on how the format by which evidence strength is communicated and the type of evidence influenced police judgments of a suspect's guilt and their accuracy in such guilt judgments.

Evidence Strength Format

The way evidence is communicated can influence how police evaluate the evidence, and ultimately how they judge suspect guilt. Statistical formats are one method to communicate evidence strength. To understand why statistical formats are useful to communicate evidence strength, consider the following example. Imagine a forensic examiner compares a fingerprint sample from a crime scene to a suspect's fingerprint. After analyzing the samples, the expert needs to convey the results to the police. To accomplish this task, forensic examiners are trying to answer the following question:

What is the probability the fingerprint sample occurred given it came from the suspect versus a randomly selected person? This probability could be presented in one of two reciprocal statistical formats: A likelihood ratio (LR) or a random match probability (RMP). These statistical classifications can convey a probability of the likelihood that one hypothesis occurs relative to another hypothesis (see Table 1 for classifications of conveying evidence strength).

Table 1

Statistical Classifications of Evidence Strength

Evidence Type	Statistical Classification	Performance index	Source
Eyewitness ID	Likelihood Ratio	$LR = \frac{\Pr(IDS\|G)}{\Pr(IDS\|I)}$	Wells & Lindsay (1980)
Forensic (e.g., DNA, fingerprint)	Likelihood Ratio	$LR = \frac{\Pr(EV\|S)}{\Pr(EV\|R)}$	Association of Forensic Science Providers (2009)
	Random Match Probability	Inverted LR	Thompson et al. (2018) Thompson & Newman (2015)

Note. The statistical classification for the eyewitness ID evidence depicts the probability the eyewitness made an identification given the suspect is guilty versus the probability the eyewitness made an identification given the suspect is innocent. The statistical classifications for forensic evidence depict the probability of the evidence occurring given the suspect left the sample (e.g., fingerprint or DNA evidence) versus the probability of the evidence given a randomly chosen person left the sample.

LRs convey the probability that the evidence occurred given one hypothesis versus another hypothesis. For example, fingerprint evidence with an LR of 10 conveys that *it is 10 times more likely the evidence occurred given the fingerprint originated from the suspect versus a randomly chosen person.* The higher the LR, the stronger the evidence is at incriminating a suspect. Alternatively, RMPs are the reciprocal of LRs in a frequency format. For example, fingerprint evidence with an RMP of 10 conveys that *one*

person in 10 would have a fingerprint consistent with the fingerprint sample from the crime scene.

Mock jurors tend to evaluate LRs more accurately than an equivalent verbal (i.e., qualitative) description of the evidence strength (e.g., *the evidence weakly supports the hypothesis of guilt*; Martire et al., 2014). However, some researchers argue LRs can be difficult for people to interpret (Martire et al., 2014; Thompson & Newman, 2015). For example, Thompson and Newman (2015) found mock jurors were more accurate at judging suspect guilt when evaluating shoeprint evidence in an RMP format than in an LR format. People tend to interpret frequencies more accurately than ratios (Gigerenzer & Hoffrage, 1995), so RMPs could be a preferred statistical classification to communicate evidence strength. Therefore, police should interpret evidence presented as an RMP more accurately than evidence presented as an LR.

However, there is mixed evidence as to whether RMPs are superior to LRs. For example, mock jurors interpreted LRs and RMPs as equally accurate when evaluating DNA evidence (Thompson & Newman, 2015) and as equally strong when evaluating blood sample evidence (Thompson et al., 2018). Sometimes participants even misunderstand RMPs, believing a lower RMP (e.g., one in 10) indicated stronger evidence than a higher RMP (e.g., one in 10,000; Thompson et al., 2018). It is also unknown how police interpret such information, even though police might encounter statistical classifications of evidence strength through forensic experts who use such classifications.

Evidence Type

Another factor that can influence guilt judgments is evidence type. People

generally perceive DNA evidence as credible (Kassin et al., 2013; Lieberman et al., 2008; Thompson & Newman, 2015), and mock jurors interpreted DNA evidence accurately regardless of whether the evidence strength was presented in an LR or RMP format (Thompson & Newman, 2015). However, not all forensic evidence is evaluated accurately. For example, mock jurors under-weighed fingerprint evidence presented as an LR such that they rendered lower guilt judgments than the evidence conveyed (Martire et al., 2014), but it is unknown whether and to what extent police interpret DNA more accurately than fingerprint evidence.

One powerful type of evidence neglected in previous research on presentation formats is eyewitness identification (ID) evidence. Eyewitness ID evidence is one of the most persuasive types of evidence, but it is not always accurate (e.g., Wells et al., 2006). Numerous studies calculate the LRs associated with different eyewitness lineup procedures to determine how accurate an eyewitness is in identifying the true suspect (e.g., Steblay et al., 2011; Wells & Lindsay, 1980), but there is a surprising lack of research examining how police investigators interpret such LRs. Police could evaluate eyewitness ID evidence less accurately than forensic evidence because people generally perceive forensic evidence as accurate and objective (Ask et al., 2008; Devine & Macken, 2016), though the extent to which is currently unknown.

Measures of Guilt

The type of guilt measure could affect how participants express their guilt judgments. For example, mock jurors were more accurate at judging guilt on a log-scale than on a statement of odds (Thompson & Newman, 2015). A log-scale is a 17-point scale from “certain to be guilty” to “impossible to be guilty”, where each item is equal to

a scale of log odds (e.g., about 99 chances in 100 that the suspect is guilty). A statement of odds measure asks participants to fill in the following sentence: based on the evidence, I believe it is ___ times more likely that the suspect is guilty than not guilty. Although a log-scale is naturally comparable to LRs and RMPs and it aids in expressing high and low probabilities of guilt, it does not allow for expressing variability in guilt judgments closer to the mid-point if a participant had weak guilt judgments. For example, participants who weakly believed suspect guilt might have judgments that would align best with a 6 in 10 chance the suspect was guilty, but this log-scale cannot account for weak guilt judgments. Thus, a log-scale could constrain participants' responses. A measure of guilt on a percent scale (0% – 100%) allows for expressing guilt judgments closer to the midpoints of the scale, but it is unknown how people translate statistical formats of evidence strength onto a percent scale.

Previous research assessed how statistical formats of evidence affected guilt judgments within the context of a criminal trial (Thompson et al., 2018; Thompson & Newman, 2015), and evaluations of evidence differed between the context of a trial or an investigation (Sommers & Douglass, 2007), but it is unknown how different formats of evidence strength (i.e., LRs and RMPs) influence the extent to which police are accurate in evaluating evidence to judge a suspect's guilt. Additionally, there is mixed research as to whether participants understand RMPs more easily than LRs. Previous research has also not examined the effects of evidence strength format on guilt judgments for eyewitness ID evidence. The proposed research will address these gaps by (1) recruiting a police sample, (2) comparing evaluations of DNA evidence to fingerprint evidence, (3) comparing evaluations of eyewitness ID evidence to forensic evidence, (4) examining

evaluations of eyewitness ID evidence in statistical formats, (5) comparing log-scale guilt measures to percentage scale guilt measures to examine how different measures capture self-reported guilt judgments.

Study Overview

Design

We randomly assigned participants to a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness ID) between-subjects factorial design. The evidence strength format factor manipulated whether participants read about moderately incriminating evidence (i.e., an LR or RMP of 100) in the LR format or the RMP format. An LR of 100 is considered to communicate moderately incriminating evidence by the Association of Forensic Science Providers (2009; see Table 2) and it has been used to manipulate moderately incriminating evidence in past research (Thompson & Newman, 2015). We chose moderately incriminating evidence for the level of evidence strength because strong evidence could be subject to ceiling effects (e.g., Thompson et al., 2018); therefore, this level allowed for variability in responses. Participants in the neutral condition read that the evidence was inconclusive and was designed to act as a control. The evidence type factor manipulated whether participants read about DNA, fingerprint, or eyewitness ID evidence.

Table 2
Descriptions of Evidence Strength for Likelihood Ratios

Value of LR	Verbal Equivalent
>1-10	Weak support for proposition
10-100	Moderate support for proposition
100-1000	Moderately strong support for proposition
1000-10,000	Strong support for proposition
10,000-1,000,000	Very strong support for proposition
>1,000,000	Extremely strong support for proposition

Research Questions and Hypotheses

This study aims to answer three primary research questions: 1) How does evidence strength format affect participants' accuracy in judging suspect guilt? 2) How does evidence type affect participants' accuracy in judging suspect guilt? 3) To what extent do participants' guilt judgments differ by guilt measure type?

Based on previous research, we offer two main hypotheses concerning the effects of evidence strength format and evidence type on accuracy:

1) Police participants who read about evidence in an RMP format will be more accurate in their guilt judgments than police participants who read about evidence in an LR format.
2) Police participants will be less accurate in their evaluations of eyewitness ID evidence when compared to fingerprint and DNA evidence, as well as less accurate in their evaluations of fingerprint evidence when compared to DNA evidence.

Method

Participants

We recruited U.S. police officers using CloudResearch, an online platform that integrates with Amazon Mechanical Turk to recruit high-quality participants (Litman et

al., 2017). We aimed to recruit at least 196 participants based on an a priori power analysis that assumed a medium effect ($f = 0.25$), a power of $1 - \beta = 0.80$, and a Type I error rate $\alpha = 0.05$. CloudResearch replaced participants who did not pass the occupation screening question (i.e., answer they were employed in law enforcement), failed the attention check, or failed the manipulation memory checks until they recruited a sample of at least 200 high-quality police participants. In total, 209 police completed the study and self-identified their gender as 73% men, 23% women, and their race/ethnicity as 72% Caucasians, 14% African American, 11% Latino/a, 2% Asian, 1% Native American, and 0.5% multi-ethnic. Participants were, on average, 37.2 years old (*SD* = 9.4; range = 18-68).

On average, participants had 8.9 years of law enforcement experience and self-identified their department as 70% local, 21% state, 4% federal, and 5% other. Participants self-reported their ranking as 18% patrol officer, 18% police officer, 17% Detective/Investigator, 8% Lieutenant, 5% Sergeant, 4% Sheriff, 2% Deputy Chief, 1% Chief of Police, and 9% other.

Materials

Crime Scenario

Participants read a prompt asking them to imagine that they are a police officer who is investigating a murder. Participants read there was a possible suspect but based on the current information there was only one chance in two (fifty-fifty chance) that the suspect is guilty.

Evidence Type

The evidence type manipulation varied whether the evidence discovered in the

case was *DNA*, *fingerprint*, or *eyewitness ID* (see Appendix). In previous research, probability values typically used for eyewitness ID evidence represented how much more likely it was that the eyewitness made an identification given the suspect is the culprit versus the suspect is innocent (e.g., Wells & Lindsay, 1980). This format of evidence provides support for whether a particular suspect committed a crime, unlike DNA and forensic evidence. DNA evidence and fingerprint evidence often only provide support as to whether a suspect was linked to a crime scene, not whether the suspect is the culprit who *committed* the crime. In other words, eyewitness ID evidence strength typically conveys the probability of committing a crime, whereas DNA and forensic evidence typically conveys the probability the suspect was at the crime scene. Thus, this study instead presented LRs and RMPs for the eyewitness ID evidence as *how much more likely the eyewitness ID was to have occurred if the suspect was at the crime scene than if the suspect was a randomly chosen person*. This presentation will avoid the potential confound that one type of evidence is directly suggesting a probability of committing a crime when the other two types of evidence can only link a suspect to a crime scene.

Evidence Strength Format

The evidence format manipulation was adapted using language from Martire and colleagues (2014), Thompson and Newman (2015), and Wells and Lindsay (1980; see Appendix). Participants either read that the evidence strength in an LR of 100 (*LR format*), as a one in 100 chance to have occurred (*RMP format*), or as inconclusive (*neutral*).

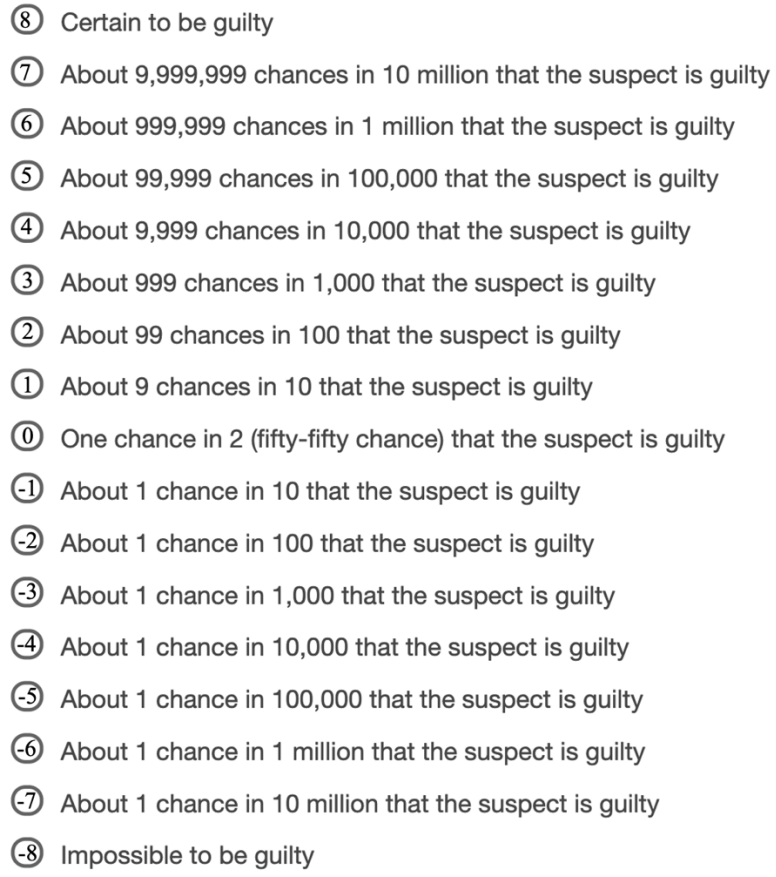
Log-scale Guilt Evaluations

We assessed participants' judgments of suspect guilt by using a log-scale. This

guilt evaluation measure assessed participants' guilt judgments using a 17-point log-scale after reading about the evidence (Martire et al., 2014; Thompson & Newman, 2015; see Figure 1 for the full scale). Except for the endpoints (*certain to be guilty; impossible to be guilty*), the intervals between the points are approximately equal on a scale of log odds.

Figure 1

Log-scale Guilt Measure



Note. Participants filled in the following statement: "Considering the information in the case, I believe the chances the suspect is guilty are___." The numbers in the response bubbles were not present to participants, rather their purpose here is to aid in readers' understanding of the calculation of the accuracy measure.

A log-scale guilt measure is the most natural to compare with LR or RMP, as they

are both on the same numeric scale. Other types of measures, such as Likert-type measures, would not be as naturally comparable to LRs or RMPs because these measures are numerically incomparable. A second justification for using a log-scale is that participants tend to be more accurate in their subjective evidence strength judgments on a log-scale than on a scale using a statement of odds (i.e., based on the evidence, I believe it is ___ times more likely that the suspect is guilty than not guilty; Thompson & Newman, 2015).

Participants read and filled in the following sentence with their perceived chances that the suspect was guilty: "Considering the information in the case, I believe the chances the suspect is guilty are___." For example, participants could have answered there are "*about 9,999,999 chances in 10 million that the suspect is guilty*" if they believed the evidence strongly indicated guilt, or they could have answered there are "*about 1 chance in 10 million that the suspect is guilty*" if they believed the evidence strongly indicated innocence (Martire et al., 2014; Thompson & Newman, 2015).

Instructions for interpreting the log-scale were based on instructions from Thompson and colleagues (2013). These instructions were piloted by recruiting participants ($N = 59$) from Amazon Mechanical Turk (MTurk) to ensure participants interpreted the log-scale correctly, that is, that participants interpreted statements on the top half of the scale as indicative of guilt (e.g., *about 9,999,999 chances in 10 million that the suspect is guilty)* and statements on the bottom half of the scale as indicative of innocence (e.g., *about 1 chance in 10 million that the suspect is guilty*). See Appendix for these instruction materials.

Percentage Scale Guilt Evaluation

Participants' judgments of suspect guilt were also assessed on a percentage scale. Although the log-scale guilt judgment most naturally aligns with LR or RMP, participants' answers could be constrained in the log-scale format. Specifically, the log-scale measure of guilt limits responses to the highest probabilities of guilt or innocence (i.e., the top and bottom 10%). Therefore, participants answered their perceived likelihood that the suspect committed murder from (*0% likely*) to 100 (*100% likely*) on a slider scale that has more variability near the mid-point (see Appendix).

Accuracy Measure

We created an accuracy measure from our log-scale guilt measure based on analyses from Thompson and Newman (2015). This accuracy score is calculated from the deviation between how participants should respond compared to their actual responses. Specifically, a *subjective evidence strength* score was calculated for each participant based on which item they chose on the log-scale. Positive values refer to items above the mid-point and negative values refer to items below the mid-point (see Figure 1). Next, we subtracted how participants should have responded (based on an LR or RMP of 100 or 1 for the neutral condition) from participants' subjective evidence strength score to create an *accuracy* measure. For example, if participants in the experimental groups chose "*about 999 chances in 1000 that the suspect is guilty*" on the log-scale, their *subjective evidence strength* score is 3. However, they should have chosen "*about 99 chances in 100 that the suspect is guilty*" because this item statistically corresponds with an LR and RMP of 100. This item on the scale corresponds to a score of 2. Thus, their accuracy score would be 1 (3-2 = 1). Positive numbers indicated the participant judged the suspect as

more guilty than they should and negative numbers indicated the participant judged the suspect as more innocent than they should. An accuracy score of 0 indicated perfect accuracy.

Suspicion, Attention, and Manipulation Check

Participants indicated what they knew about the study, if they had any suspicions, and described their knowledge of the study or suspicions (see Appendix). Participants also self-reported their attention by answering whether they paid attention and if they read the full case scenario. The survey did not permit participants who self-reported they did not pay attention to continue. To assess participants' memory of the evidence type manipulation, we asked whether they read about DNA, fingerprint, or eyewitness ID evidence. To assess participants' memory of the evidence format manipulation, we asked whether the evidence they read was in the LR format, RMP format, or inconclusive. The survey logic did not permit participants who failed to correctly report their memory of the manipulations to continue taking the survey. Finally, we embedded a visually hidden question designed so human responders cannot see the question, but bots could detect the question and answer it. Qualtrics redirected participants who answered the bot screener question out of the survey so CloudResearch could replace them with new participants.

Demographic Questionnaire

At the beginning of the survey, participants reported their occupation from a list of 20 work industries (e.g., real estate, retail) to identify participants who work in law enforcement. CloudResearch did not permit participants who answered any item except "law enforcement" on the occupation screener question to continue taking the survey and replaced them with new participants. At the end of the survey, participants answered

questions regarding their number of years of law enforcement experience, the type of agency they currently work for (i.e., local, state, or federal), and their current department ranking. Participants also answered a basic demographic questionnaire assessing age, gender, race and ethnicity, education, and political identity (see Appendix). Finally, Qualtrics redirected participants back to CloudResearch for compensation.

Procedure

CloudResearch recruited participants to complete the study via Qualtrics. After reading an informed consent sheet, participants read a prompt that asked them to imagine they were investigating a murder case and there was only a one chance in two that the suspect was guilty. Next, participants read a prompt that new evidence in the case emerged. Qualtrics randomly assigned participants to one of the nine experimental groups based on evidence format and evidence type. After reading about the randomly assigned piece of evidence, participants answered their judgments on guilt on the log-scale and the percent guilt scale. On the next page, participants answered the manipulation check, suspicion check. On the last page, participants answered the demographics questionnaire.

Results

We first examined how accurate the police were in their judgments of suspect guilt on the log-scale guilt measure to test our hypotheses regarding the main effects of evidence strength format and evidence type on accuracy. Next, we analyzed how evidence strength format and evidence type affected guilt judgments. Separately analyzing guilt judgments before converting the log-scale to the accuracy measure can aid in parsing out the extent to which police were judging the suspect to be guilty and whether police were judging suspect guilt similarly on the log-scale as the percentage

guilt scale. We only had predictions regarding the *accuracy* scores, not the raw log-scale or percent guilt measures, so analyses on these guilt measures were exploratory.

Accuracy

We first analyzed a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-groups factorial ANOVA on the accuracy measure to test our hypotheses regarding the effects of evidence strength format and evidence type on police's accuracy in judging suspect guilt (see Figure 2). There was no main effect of evidence type (p = 0.53), but there was a main effect of evidence strength format, $F(2, 200) = 11.64$, $p < .001$, $\eta_p^2 = 0.10$. Surprisingly, post hoc comparisons using Tukey's Honest Significant Difference (HSD) test indicated that the mean accuracy score among participants in the *RMP condition* (M = -1.82, SD = 2.66) was significantly lower (i.e., less accurate) than in the *LR condition* (M = -0.21, SD = 2.79), $t(200) = 3.46$, $p = .002$, $d = 0.58$, 95% CI [0.50, 2.67], and the *neutral condition* (M = 0.37, SD = 2.87), $t(200) = 4.66$, $p < .001$, $d = 0.80$, 95% CI [1.08, 3.29]. There was no difference between the *LR condition* and the *neutral condition* on accuracy (p = .40). Thus, our hypotheses were not substantiated. Recall that participants' guilt judgments are more accurate the closer their accuracy score is to 0. We found participants in the *RMP condition* significantly under-weighed the evidence when compared to the *LR condition,* and thus they were less accurate in judging suspect guilt.

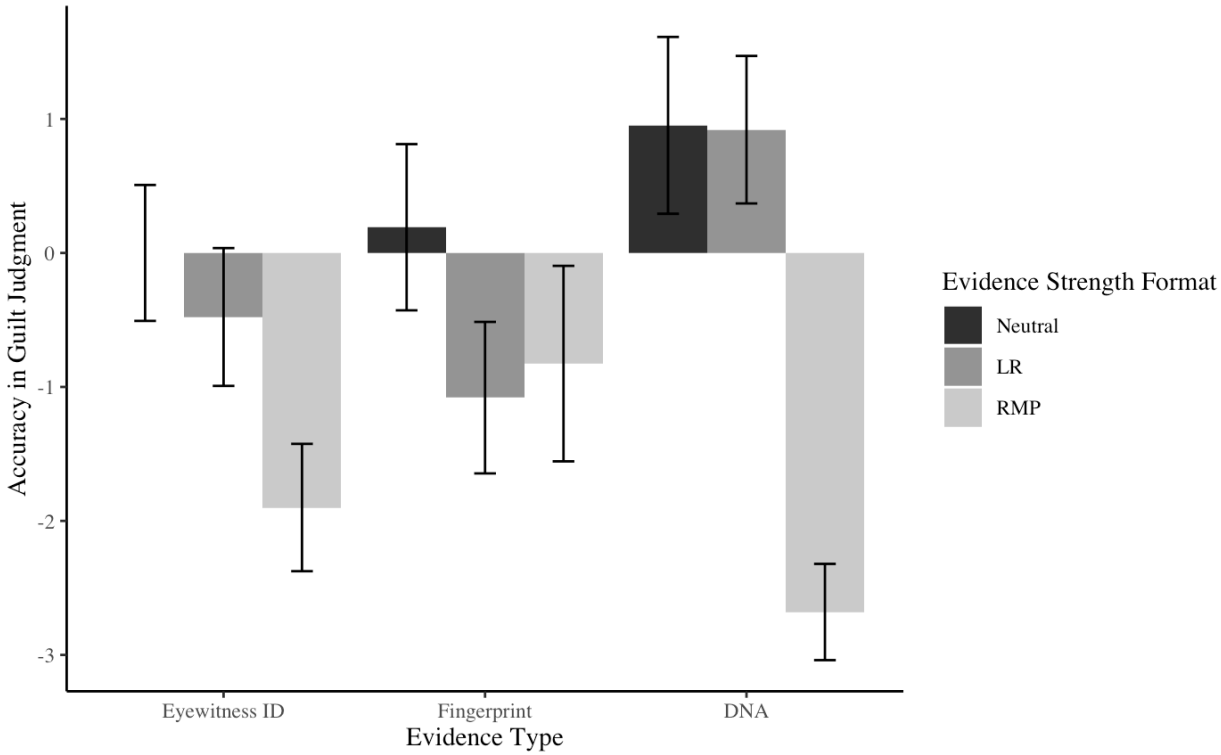
This main effect was qualified by a significant interaction between evidence strength format and evidence type, $F(4, 200) = 3.23$, $p = .014$, $\eta_p^2 = 0.06$. Post hoc comparisons using Tukey's HSD adjustments indicated participants who saw *DNA evidence* in the *RMP format* (M = -2.68, SD = 1.80) significantly under-weighed the

evidence when compared to participants who saw the *DNA evidence* in the *LR condition* (M = 0.92, SD = 2.75), $t(200) = 4.68$, $p < .0001$, $d = 1.32$, 95% CI [1.78, 5.42], and the *neutral condition* (M = 0.95, SD = 3.02), $t(200) = 4.51$, $p < .0001$, $d = 1.34$, 95% CI [1.73, 5.53]. There were no other significant effects of evidence strength format by evidence type, (ps > .06). Thus, our findings suggested that participants were more accurate at judging the suspect in the *LR condition* (compared to the *RMP condition*) when participants encountered *DNA evidence*.

We also found that participants in the *LR condition* significantly under-weighed *fingerprint evidence* (M = -1.08, SD = 2.83) when compared to *DNA evidence* (M = 0.92, SD = 2.75), $t(200) = 2.60$, $p = .027$, $d = 0.74$, 95% CI [0.18, 3.82]. Although participants' accuracy scores in the *LR condition* significantly differed depending upon whether they encountered *fingerprint* or *DNA evidence*, their mean scores suggest they were similarly accurate in terms of an absolute value because both groups deviated from accuracy, on average, by approximately one accuracy point but in opposite directions. There were no other significant interaction effects in the *neutral*, *LR*, or *RMP* conditions (ps > .05).

Figure 2

Police Accuracy in Judging Suspect Guilt



Note. This graph shows the mean accuracy scores by evidence strength format and evidence type with standard error bars. Higher mean scores suggest police were over-weighing the evidence (i.e., interpreting the evidence as more incriminating than they should), whereas lower mean scores suggest police were under-weighing the evidence (i.e., interpreting it as less incriminating than they should). A score of 0 suggests perfect accuracy. Police who encountered eyewitness ID evidence condition in the neutral condition rendered a mean accuracy of 0, thus there is no bar graph for this condition but there are still standard error bars.

Guilt Judgments

One of our research questions concerned how participants reported their judgments of suspect guilt on two different measures: a log-scale guilt measure and a percent guilt measure. Therefore, we analyzed participants' raw log-scale guilt judgments in addition to their accuracy scores (reported above) to compare trends in guilt judgments between measures.

Log-Scale Guilt Judgments

We conducted a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-groups factorial ANOVA on log-scale guilt judgments to examine how evidence strength format and evidence type affect participants' raw guilt judgments on the log-scale (see Figure 3). Although we did not find a significant main effect of evidence type ($p = 0.48$), we did find a significant main effect of evidence strength format, $F(2, 200) = 7.24$, $p < .001$, $\eta_p^2 = 0.07$. Post hoc comparisons using Tukey's HSD adjustments indicated that participants' mean log-scale guilt judgments in the *LR condition* ($M = 1.79$, $SD = 2.79$) was significantly higher than in the *RMP condition* ($M = 0.20$, $SD = 2.66$), $t(200) = 3.46$, $p = .002$, $d = 0.58$, 95% CI [0.50, 2.67], and the *neutral condition* ($M = 0.37$, $SD = 2.87$), $t(200) = 3.06$, $p = .007$, $d = 0.52$, 95% CI [0.32, 2.49]. There was no difference between the *RMP condition* and the *neutral condition* on log-scale guilt judgments ($p = .92$). Thus, police judged evidence presented as an RMP similarly to neutral evidence despite being the same evidence strength as evidence presented as an LR, further confirming that police were under-weighing evidence presented in the RMP format.

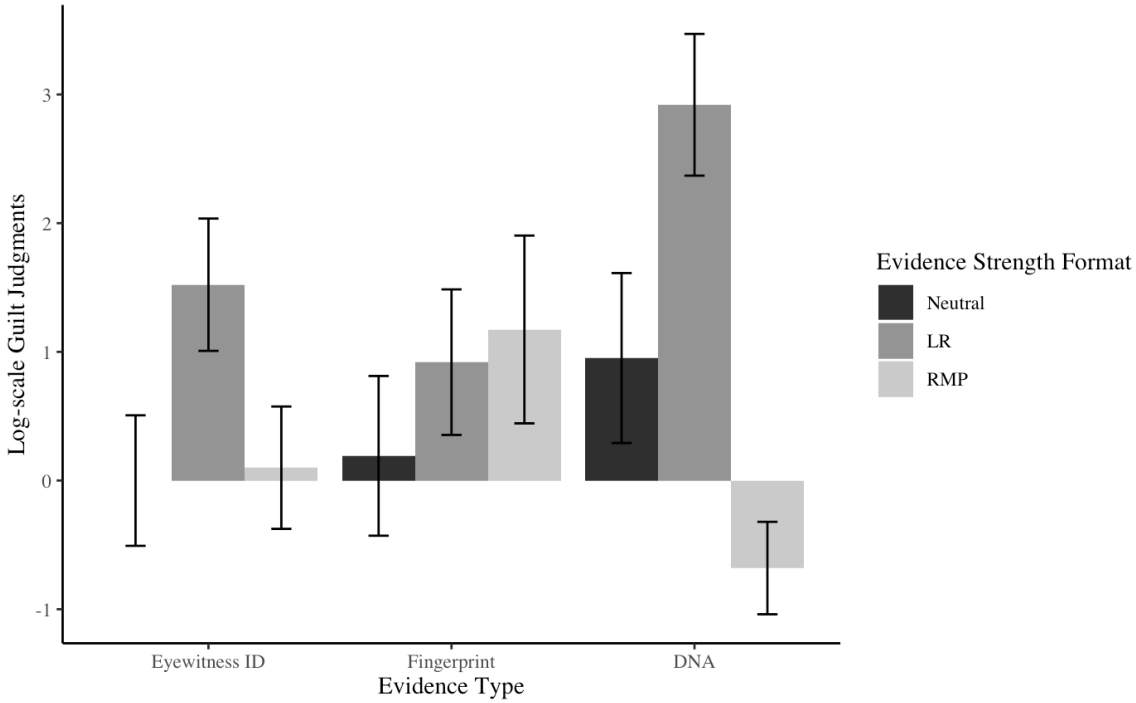
However, this main effect was qualified by a significant interaction between evidence strength format and evidence type, $F(4, 200) = 3.23$, $p = .014$, $\eta_p^2 = 0.06$. Post hoc comparisons using Tukey's HSD adjustments indicated participants who saw *DNA evidence* in the *LR condition* ($M = 2.92$, $SD = 2.75$) rated the suspect as significantly more guilty than participants in the *RMP condition* ($M = -0.68$, $SD = 1.79$), $t(200) = 4.68$, $p = < .001$, $d = 1.32$, 95% CI [1.78, 5.42], and the *neutral condition* ($M = 0.95$, $SD = 3.02$), $t(200) = 2.44$, $p = .041$, $d = 0.72$, 95% CI [0.00, 3.87]. There were no other

interaction effects of evidence strength format by evidence type (*p*s > .10). As expected, participants interpreted the *DNA evidence* in the *LR condition* as more incriminating than participants in the *neutral condition.* However, this trend did not hold among participants who encountered *DNA evidence* in the *RMP* condition because they surprisingly judged there was a less than fifty-fifty chance the suspect was guilty.

We also found that participants judged *DNA evidence* (*M* = 2.92, *SD* = 2.75) as significantly more incriminating than *fingerprint evidence* (*M* = 0.92, *SD* = 2.83) in the *LR condition* $t(200) = 2.60$, $p = .027$, $d = 0.74$, 95% CI [0.18, 3.82]. This effect suggests participants perceived DNA evidence as more incriminating than fingerprint evidence, despite conveying the same evidence strength. There were no other significant interaction effects in the *neutral*, *LR*, or *RMP* conditions (*p*s > .05).

Figure 3

Police Log-scale Guilt Judgments



Note. This graph depicts participants' mean guilt judgments on the log-scale measure with standard error bars. Each number on the log-scale (y-axis) corresponds to a certain chance of guilt (see Figure 1). The points are approximately equal on a scale of log odds. Positive values indicate higher chances of guilt, whereas negative values indicate lower chances of guilt. A score of 0 suggests participants rated there was 1 chance in 2 (fifty-fifty chance) that the suspect was guilty. Participants in the neutral eyewitness ID evidence condition had a mean score of 0 and thus there is no bar graph for this condition but there is still a standard error bar.

Percent Guilt Judgments

We conducted a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification) between-groups factorial ANOVA to examine how evidence strength format and evidence type affect participants' guilt judgments on the percent guilt scale (see Figure 4). Similar to the log-scale guilt judgments, we did not find a main effect of evidence type but we did find a main effect of evidence strength format, $F(2, 200) = 13.31$, $p < .001$, $\eta_p^2 = 0.11$ such that participants in the *LR condition* ($M = 74.47\%$, $SD = 22.67\%$) judged the suspect as significantly guiltier than participants in the *RMP condition* ($M = 57.0\%$, $SD = 29.99\%$), $t(200) = 4.14$, $p = .002$, $d = 0.70$, 95% CI [7.46, 27.29], and the *neutral condition* ($M = 55.30\%$, $SD = 22.94\%$), $t(200) = 4.53$, $p < .001$, $d = 0.77$, 95% CI [9.09, 28.93]. There was no difference between the *RMP condition* and the *neutral condition* ($p = .92$). Again, police judged evidence presented as an RMP similarly to neutral evidence despite conveying the same evidence strength as LRs.

There was also a significant interaction effect that qualified the main effect of evidence strength format, $F(4, 200) = 2.63$, $p = .035$, $\eta_p^2 = 0.06$. Post hoc comparisons using Tukey's HSD adjustments indicated participants who saw *DNA evidence* in the *LR format* ($M = 83.47\%$, $SD = 21.58\%$) rated the suspect as significantly more guilty than

participants who saw *DNA evidence* in the *RMP format* (M = 51.12%, SD = 29.84%), $t(200) = 4.61$, $p = .0002$, $d = 1.30$, 95% CI [15.77, 48.95]. There were no other significant differences in evidence strength format among participants who saw *DNA evidence* (ps > 0.06). This interaction effect is in the same direction as the log-scale guilt judgment. Thus, participants on both the log-scale guilt measure and the percent scale guilt measure rated *DNA evidence* presented in an *LR format* as more incriminating than *DNA evidence* presented in an *RMP format*.

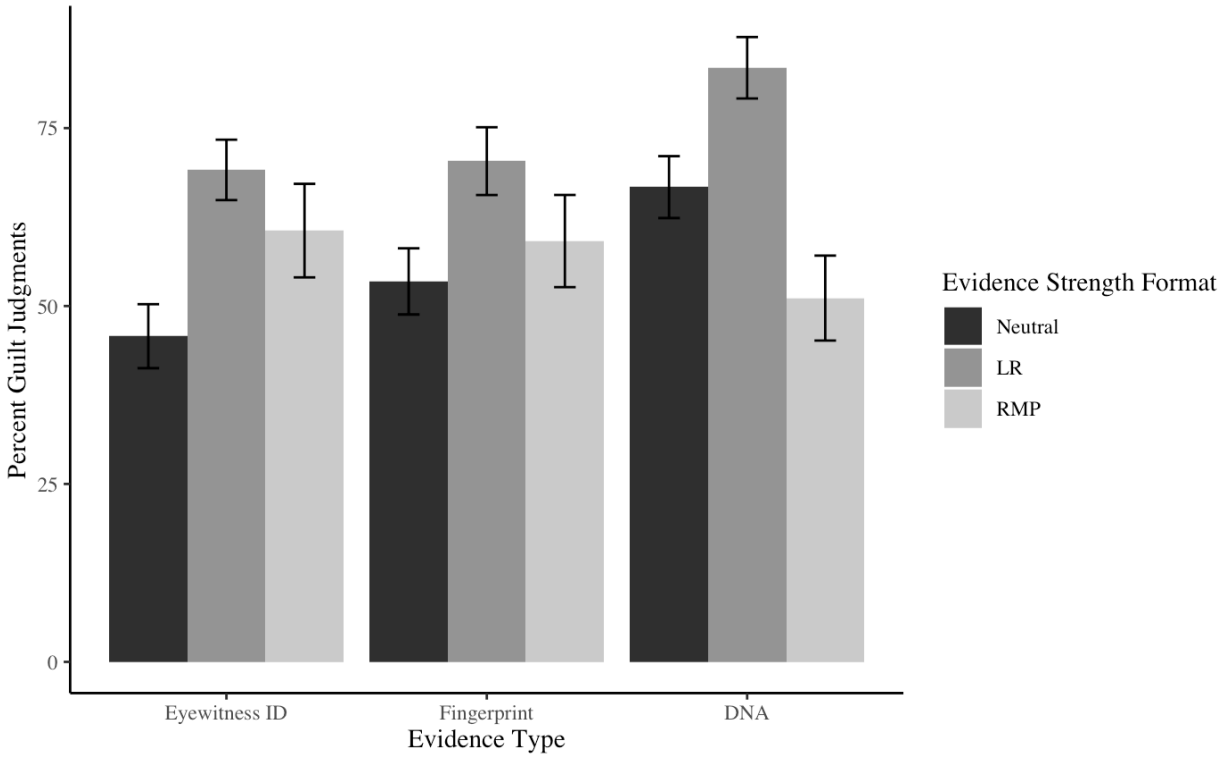
Participants who encountered *eyewitness ID evidence* interpreted evidence in the *LR condition* (M = 69.13%, SD = 20.35%) as more incriminating than evidence in the *neutral condition*, (M = 45.76%, SD = 20.61%), $t(200) = 3.12$, $p = .006$, $d = 0.94$, 95% CI [5.67, 41.07]. Similarly, participants who encountered *fingerprint evidence* interpreted evidence presented as an *LR* (M = 70.36%, SD = 23.80%) as more incriminating than evidence in the *neutral condition*, (M = 53.46%, SD = 23.73%), $t(200) = 2.43$, $p = .042$, $d = 0.68$, 95% CI [0.47, 33.33]. Both effects are expected, as the LR evidence was designed to be moderately incriminating whereas the neutral evidence was not. There were no other significant effects of evidence strength format among participants who saw *eyewitness ID evidence* or *fingerprint evidence* (ps > .13).

Unlike participants' judgments on the log-scale measure, there was an effect of evidence type by evidence format on the percent guilt measure. Participants in the *neutral condition* judged the suspect as significantly less guilty when they encountered *eyewitness ID* (M = 45.76%, SD = 20.60%) than *DNA evidence* (M = 66.71%, SD = 19.94%), $t(200) = 2.73$, $p = .019$, $d = 0.84$, 95% CI [2.85, 39.06]. Participants interpreted neutral DNA evidence as significantly more incriminating than neutral eyewitness ID

evidence on the percent guilt scale, suggesting that police perceived neutral DNA evidence as more indicative of suspect guilt when compared to neutral eyewitness ID evidence.

Figure 4

Police Percent Guilt Judgments



Note. This graph depicts police's guilt judgments on the percent guilt scale with standard error bars. Higher values indicate a greater likelihood of guilt, whereas lower values indicate a lower likelihood of guilt.

Discussion

Police need to accurately interpret evidence presented in statistical formats (e.g., LRs and RMPs). Forensic experts often use such statistical formats to convey the strength of forensic evidence (e.g., Martire et al., 2014), and researchers also use these formats to quantify eyewitness evidence strength (e.g., Wells & Lindsay, 1980). This research

examined how the statistical formats of evidence strength and evidence type influenced police accuracy in judging suspect guilt and the extent to which their guilt judgments differed by separate guilt measures. We first discuss our main effects, as these were our predicted effects, and then discuss how these main effects were qualified by our exploratory interaction analyses. Next, we discuss implications for psychological research and implications of our overall findings for policy. Last, we review the limitations of the present work.

Main Effects of Evidence Strength Format on Guilt Judgments and Accuracy

Police in the LR condition judged the suspect to be more guilty than participants in the neutral condition. This effect is expected because the LR condition was designed to communicate moderately incriminating evidence, whereas the neutral condition was designed to be inconclusive and thus offered no new information regarding the suspect's guilt status. Further, mean scores for police in the LR condition were close to 2 (M = 1.79), whereas mean scores for the police in the neutral evidence condition were close to 0 (M = 0.37), suggesting police were mostly accurate in interpreting these formats. Indeed, mean scores on the accuracy measure were close to 0 and not significantly different when comparing the LR condition (M = -0.21) to the neutral condition (0.37).

However, police judgments of suspect guilt in the RMP condition did not follow our predictions that police would more accurately judge suspect guilt when we presented evidence strength as an RMP than an LR because people tend to interpret frequencies more accurately than ratios (Gigerenzer & Hoffrage, 1995). We instead found police more accurately judged suspect guilt when we presented evidence strength in an LR format than an RMP format. Further, participants in the RMP condition and the neutral

condition judged the suspect as equally guilty on the log-scale and the percent scale, suggesting police were more conservative in their guilt judgments in the RMP condition than the LR condition. That is, police interpreted RMPs as significantly less incriminating when compared to LR, despite both communicating the same evidence strength in different formats.

One possible reason that RMPs had a similar effect on guilt judgment as neutral evidence could be that police might have had exposure to LRs before, but perhaps they have not had exposure to RMPs. Forensic experts already use LRs to convey forensic evidence strength (Association of Forensic Science Providers, 2009), so it is possible police have had prior exposure to evidence presented in an LR format. Police with little exposure to RMPs might interpret the evidence as offering no new information if they are unfamiliar with the format. Future research could add additional questions asking which formats of evidence police had previous exposure to explore the possibility that previous exposure to different statistical formats of evidence strength is affecting responses.

Another possible reason for the misinterpretation of RMPs could be explained by Thompson and colleagues (2018). These authors found some participants misinterpreted RMPs as statements regarding how many possible suspects there were, rather than a statement regarding probability from a sample of random people. In other words, some participants concluded evidence is stronger when the match frequency is high (e.g., one in 10 or lower), rather than concluding the evidence is weaker when the match frequency is high. For example, rather than interpreting “the forensic expert concluded one in 100 people would have a fingerprint that is consistent with the fingerprint sample found at the crime scene” as moderately incriminating, they might have instead interpreted this

evidence as a “one in 100 chance this person is the suspect.” It is possible police interpreted the evidence as very weakly incriminating if they were interpreting evidence presented in the RMP condition as there was a “one in 100 chance this person is the suspect.”

We did attempt to clearly state the probability for the RMP condition was referencing a group of randomly chosen people, rather than a group of suspects, by changing the wording of the evidence from Thompson and colleagues (2018). Specifically, Thompson and colleagues (2018) stated “given the size and quality of the crime scene print I would expect about one person in 1000 to have a fingerprint similar enough to be indistinguishable from it.” In our attempt to clarify the comparison group for evidence strength presented as an RMP, we worded the RMP condition for fingerprint evidence as follows: “The forensic expert concluded one in 100 people would have a fingerprint that is consistent with the fingerprint sample found at the crime scene. That means that there is one chance in 100 of finding a consistent fingerprint pattern from a randomly chosen person.” Future research should further examine how police are interpreting RMPs to identify the discrepancy between the intended evidence strength and participants’ subjective evaluations of the evidence strength.

Finally, the legal context in which we conveyed the evidence might be one explanation as to why our findings on the effects of evidence strength format differed from past research. Previous research examined evaluations of evidence presented statistical formats within the context of a trial (Martire et al., 2014; Martire et al., 2013; Thompson & Newman, 2015), whereas we examined these evaluations within the context of an investigation. For example, alibi evidence led to different guilt judgments

depending on whether the evidence was within the context of an investigation or a trial (Sommers & Douglass, 2007). Future research could examine if the effect of evidence strength format on judging suspect guilt differs between whether the evidence is presented during an investigation or a trial.

Lack of Main Effects of Evidence Type on Guilt Judgments and Accuracy

Previous psychological research suggests that people tend to perceive DNA evidence as very credible (Kassin et al., 2013; Lieberman et al., 2008; Thompson & Newman, 2015) and as the "gold standard" of evidence (Ask et al., 2008; Kassin et al., 2013). Although people do not perceive fingerprint evidence to be as credible as DNA evidence, the criminal justice system has used it for over 100 years and people tend to weigh it heavily in their guilt judgments (Garrett et al., 2020; Garrett & Mitchell, 2013). People tend to also place great weight on eyewitness ID evidence when making decisions, but eyewitness testimony is increasingly recognized as fallible and as less reliable than forensic evidence, whereas forensic evidence is generally perceived as accurate and objective (Ask et al., 2008; Devine & Macken, 2016). Thus, we predicted that police would judge suspect guilt after encountering DNA evidence most accurately, then fingerprint evidence, then eyewitness ID evidence. However, this prediction was unsubstantiated.

Our data did not support the hypothesis that evidence type affects guilt judgments or accuracy. The means were trending in the predicted direction, that is, police interpreted *DNA* (M = -0.34) more accurately than *fingerprint evidence* (M = -0.55), and both *DNA* and *fingerprint evidence* more accurately than *eyewitness ID evidence* (M = -0.77). However, this trend was not statistically significant. We found a very small effect size for

this non-significant effect ($\eta_p^2 = 0.003$), and thus a larger sample size might be needed to detect whether there are differences in police guilt judgments by evidence type. Further, the means of the accuracy score were all close to 0, suggesting police judged suspect guilt mostly accurately regardless of evidence type, although they did slightly under-weigh the evidence on average.

One possible explanation as to why police were somewhat accurate in their guilt judgments by evidence type could be that people tend to interpret numerical formats of evidence strength more accurately than their verbal equivalents (Martire et al., 2014). Perhaps police interpreted this evidence accurately because it was in a numerical format. For example, consider the study by Ask and colleagues (2008) that found a police trainees' interpretations of a witness statement were more influenced by contextual variables compared to DNA evidence. It is possible that communicating the evidence in a numerical format in our study led police to similarly interpret all types of evidence when they are similar strengths.

Interaction Between Evidence Strength Format and Evidence Type on Accuracy

Although we initially only predicted main effects resulting from our manipulated factors, our exploratory analyses revealed an interaction between evidence strength format and evidence type on accuracy and on both guilt measures. First, we found an effect of evidence strength format on the log-scale and percent guilt measure among police participants who were in the DNA evidence condition. Participants in the DNA evidence condition reported the suspect was guiltier in the LR condition than in the RMP condition on both their log-scale and percent guilt judgments, whereas they interpreted evidence presented in the RMP condition similarly to evidence presented in the neutral

condition. The accuracy measure further revealed that participants were over-weighing the DNA evidence in the LR condition by almost one (M = 0.92) accuracy point and under-weighing evidence in the RMP condition by almost 3 (M = -2.68) accuracy points. Still, participants in the DNA evidence condition who read the evidence in the LR format were significantly more accurate in their guilt judgments than participants who read the evidence in the RMP format.

Thompson and Newman (2015) found an interaction between evidence strength format and evidence type, albeit in the opposite direction. Specifically, they found mock jurors rendered more accurate guilt judgments for shoeprint evidence presented in the RMP format versus the LR format and no difference between evidence strength formats for the DNA evidence. Perhaps the difference in our sample from Thompson and Newman (2015) could explain this discrepancy. A published survey, though not peer-reviewed, found police and legal professionals perceived DNA evidence as more reliable, trustworthy, and influential than laypeople (Scudder et al., 2020). Police might have more experience with DNA as the "gold standard" of evidence compared to other types of evidence, which could explain why this type of evidence produced the largest shifts in guilt judgments compared to the other types of evidence. If the general trend is for police to interpret RMPs as less incriminating compared to LR, then this effect could be exacerbated in the DNA evidence condition due to it being considered the "gold standard." Scudder and colleagues (2020) did not separate police officers from other legal professionals or define legal professionals, so future research should consider examining how police perceive various types of evidence as compared to laypeople to provide understanding as to how police perceive the reliability of different evidence types. Future

research should attempt to replicate our findings to ensure this effect is not simply an artifact of our study, as well as recruit a layperson sample to offer a direct comparison to past mock juror research.

It is also possible that DNA evidence presented as an RMP is subject to the weak evidence effect. The weak evidence effect occurs when evidence that weakly supports propositions of guilt leads to a decrease in beliefs of guilt rather than a small increase in beliefs of guilt (e.g., Martire et al 2014; Martire et al., 2013). Indeed, police depicted there was less than a fifty-fifty chance the suspect was guilty when they encountered DNA evidence in the RMP condition (log-scale mean guilt = -0.68). Because police initially read there was only "one chance in two (a fifty-fifty chance) that the suspect is guilty", this response suggests a decrease in guilt judgments. However, there is only partial support for the weak evidence effect because participants interpreted the DNA evidence in the RMP condition as slightly above a fifty-fifty chance of guilt on the percent guilt scale (M = 51.12%). Thus, our findings offer mixed support for whether DNA evidence is subject to the weak evidence effect.

As expected, we found participants judged suspects as guiltier when they encountered evidence in the LR condition than the neutral condition for both eyewitness and fingerprint evidence on the percent guilt scale. This effect is expected because the LR evidence was designed to be more incriminating than neutral evidence. However, we did not find this effect when police judged suspect guilt on the log-scale, which is surprising because it implies participants interpreted both evidence strength formats as similarly incriminating for eyewitness and fingerprint evidence when compared to neutral evidence. Perhaps the lack of variability near the mid-point of the log-scale contributed to

this difference, as police could express greater variability on weaker guilt judgments on the percent guilt measure than on the log-scale guilt measure.

We also found a difference between evidence type that was contingent upon evidence strength format. Participants in the LR condition judged the suspect as guiltier on the log-scale when they encountered DNA evidence than fingerprint evidence. The accuracy measure further clarified that participants who encountered DNA evidence presented as an LR were over-weighing the evidence by approximately one point (M = 0.92), whereas they were under-weighing the fingerprint evidence presented as an LR by approximately one point (M = -1.08). Thus, participants were similarly accurate when judging suspect guilt after encountering evidence in the LR condition regardless of whether the evidence was DNA or fingerprint evidence, but they differed in whether they over-weighed or under-weighed the evidence. Our findings that police under-weighed fingerprint evidence replicate past research that found mock jurors under-weighed fingerprint evidence presented as an LR (Martire et al., 2014), although we provided the first comparison between DNA and fingerprint evidence.

Guilt Judgment Measures

We used two measures to examine police guilt judgments: log-scale and percent guilt. Our findings produced a similar pattern for main effects on both measures. First, there were no effects of evidence type on either guilt measure, except for the interaction between DNA and evidence format. Second, police in the RMP condition had similar guilt judgments as police in the neutral condition, regardless of guilt measure type. Third, police in the LR condition had significantly higher incriminating guilt judgments compared to those in the RMP and neutral condition across both guilt measures.

Despite the similar main effect trends between guilt measures, the absolute numeric value differed. For example, police in the LR condition had a mean score of 1.79 which corresponds to between 1 and 2 on the log-scale (i.e., between about 9 chances in 10 that the suspect is guilty and about 99 chances in 100 that the suspect is guilty). In percentage terms, this mean log-scale score corresponds to between 90%-99% chances of guilt. However, police in the LR condition rated the likelihood of the suspect being guilty on the percent guilt measure as, on average, 74.47%. Police, therefore, were not reporting similar numerical representations of their guilt judgments on both scales.

This discrepancy between the guilt judgment measures is more pronounced when examining the interaction between evidence strength format and evidence type on the log-scale and percent guilt measures. As previously discussed, police who encountered DNA evidence in the RMP condition perceived the chances of suspect guilt as below fifty-fifty on the log-scale, whereas they perceived the likelihood of suspect guilt as slightly above 50% on the percent guilt scale.

Another discrepancy was that the pattern of results for police guilt judgments differed when comparing types of evidence across different evidence formats. On the log-scale, participants judged the suspect as more guilty after encountering DNA evidence compared to fingerprint evidence in the LR condition. On the percentage guilt scale, the only difference between evidence type by evidence strength format was that participants judged the suspect as guiltier when encountering neutral DNA evidence than neutral eyewitness ID evidence. Finally, evidence strength format had no effect on police who encountered eyewitness ID evidence or fingerprint evidence when examining log-scale guilt judgments, whereas police correctly reported LRs were more incriminating than

neutral evidence on the percent guilt judgment scale. These differing interaction effects provide further support that police are not reporting their guilt judgments similarly on both measures.

This discrepancy could leave readers wondering which measure is most accurate at measuring police's actual guilt judgments. As mentioned previously, the log-scale was limited by constraining responses to the ends of a guilt scale, whereas the percent guilt scale was limited in that it did not allow for differentiation on very high or low guilt judgments. Thus, our data suggest an important methodological implication that using different measures of suspect guilt could result in different patterns. When looking at the main effects, these measures seem to be reliable measures of guilt, as they both produced similar patterns of findings. However, these measures were not valid in numerically translating exactly how guilty police are judging a suspect to be and were not reliable when considering the pattern of results for the interaction effects.

This is not the first study to find discrepancies between measures of guilt judgments for statistical classifications of evidence strength. Thompson and Newman (2015) used a log-scale in addition to a statement of odds (e.g., based on the evidence, I believe it is ___ times more likely that the suspect is guilty than not guilty). They found police were more accurate in their judgments of defendant guilt on a log-scale than a statement of odds. Thus, based on our findings and past research (Thompson & Newman, 2015), perhaps a log-scale is the best measure to capture guilt responses when evidence strength is presented in a statistical format because it does constrain guilt judgments to the endpoint—just as these statistical representations of evidence strength often do. As it is outside of the scope of this paper to construct an instrument to measure suspect guilt,

future research should further examine what scales are most accurate for measuring police judgments of suspect guilt to ensure our tools for measuring guilt judgments are both reliable and valid. Future research should also consider using measures that consider the chances or likelihood that the suspect is innocent, either in the same measure or in a separate measure as a likelihood the suspect is guilty.

Implications for Psychology and Law Research

The present research advances the field of psychology and law in several ways. First, our work is the first to extend mock juror research on how statistical formats of evidence strength affect guilt judgments in the context of a police investigation. How police understand evidence is critical to understand, as police investigations are the first step in a criminal case. We identified several differences in our patterns of findings from mock juror research, suggesting this difference in context and sample could contribute to police understanding evidence in an investigation that is presented in a statistical format differently from laypeople responding as mock jurors in a trial.

Previously, psychological research provided mixed results as to whether LRs or RMPs were the best format for informing guilt judgments (Thompson & Newman, 2015; Thompson et al., 2018), although people tend to interpret frequencies more accurately than ratios in general (Gigerenzer & Hoffrage, 1995). Our research suggests the opposite of Gigerenzer and Hoffrage (1995), such that we found LRs are the best format to communicate evidence strength within the context of a police investigation. Interestingly, our study is the first to find LRs are better at communicating evidence strength than RMPs, as past research either found participants were more accurate at understanding RMPs compared to LRs (Thompson & Newman, 2015) or equally accurate in

understanding RMPs to LRs (Thompson et al., 2018). Perhaps the context of an investigation with police officers contributed to this difference. Regardless, our research provides a foundational understanding that police more easily interpret evidence strength when presented in an LR compared to RMP that future research can build upon.

Finally, our research examined how people evaluate eyewitness ID evidence presented in a statistical format. Researchers calculate LRs associated with different lineup procedures (e.g., Steblay et al., 2011; Wells & Lindsay, 1980), yet there is no empirical understanding of how police interpret such LRs. Overall, we found no difference in accuracy by evidence strength format when police encountered eyewitness ID evidence, although we did find differences between evidence strength formats on the percent guilt measure. Police who encountered eyewitness ID evidence judged the suspect as guiltier on the percent guilt scale when they were in the LR condition than in the neutral condition, but there was no difference when they were in the RMP condition compared to the neutral condition. This pattern of results suggests police better understand LRs than RMPs, providing a foundational understanding of how police understand statistical formats of eyewitness ID evidence strength.

Policy Implications and Recommendations

Our findings helped elucidate which formats of evidence strength are most conducive to errors during criminal investigations. Because the first step in combatting error is to identify when errors occur, our findings suggest police could benefit from additional training on RMPs. Otherwise, police seemed relatively accurate in interpreting neutral evidence and evidence presented in an LR format. Further, our interaction effects suggested even when there was an interaction between evidence strength format and

evidence type, police interpreted LRs more accurately than RMP for DNA evidence. Police were also similarly accurate when evaluating evidence presented as an LR across evidence types.

Although there are data that suggest laypeople accurately evaluate evidence presented as an RMP (e.g., Goodman, 1992; Smith et al., 1996; Thompson & Newman, 2015), our data suggest the opposite with a police sample such that police under-weigh evidence strength when presented in an RMP format, particularly with DNA evidence. Under-weighing evidence in an RMP format could be problematic in that it might lead to police not arresting guilty culprits. Rather, forensic experts and researchers should use LRs when conveying evidence strength in a statistical format to police.

There is some evidence that police do not always interpret witness evidence similarly to DNA (e.g., Ask et al., 2008). Because evidence strength in a statistical format improves accuracy (Martire et al., 2014), police could benefit from forensic examiners reporting evidence strength in a statistical format. Further, police could benefit from conveying evidence strength of other types of evidence they typically collect (e.g., eyewitness ID, confession) in statistical formats to other officers when trying to communicate evidence strength to an investigative team. LRs are used within research on eyewitness IDs (e.g., Steblay et al., 2011; Wells & Lindsay, 1980) and confession evidence (e.g., Horgan et al., 2012; Russano et al., 2005). Thus, it might be useful for police to convey such evidence in statistical formats to ensure evidence strength is interpreted similarly across different officers. Because we found that police interpreted LRs more accurately than RMPs, especially for DNA evidence, we would suggest forensic experts and researchers use LRs when statistically conveying evidence strength

during police investigations.

Limitations

This work aids in establishing patterns regarding how statistical formats of evidence strength affect police officers' psychological processing of evidence and judgments of suspect guilt, but this study only can generalize to a small piece of a complex criminal investigation. Thus, this study lacked ecological validity in a few major ways. For example, police participants knew the scenario they read did not depict a real crime or suspect. Police participants also only spent approximately 10 minutes evaluating the case and judging suspect guilt, whereas an actual investigation could last anywhere between hours and years. Police in an actual investigation could also realistically encounter more than one piece of evidence during an investigation, whereas in our study we limited the scenario to one piece of evidence for experimental purposes. The written format of our study might not be generalizable to an actual police investigation that involves communicating with a team of officers or working with tangible evidence (e.g., seeing the physical fingerprint evidence). Thus, this study lacks verisimilitude and consequentiality. Some authors suggest limitations of verisimilitude and consequentiality are minor for jury decision-making research (e.g., Bornstein et al., 2017; Bornstein & McCabe, 2004), so perhaps these limitations are minor when applied to a police investigation as well. Future research could examine the extent to which limitations of verisimilitude and consequentiality affect decision-making when applied to police investigations. Future research could also consider conducting more realistic research, such as by including multiple pieces of evidence.

Another limitation is related to a lack of consideration regarding the social

situation between researcher and participant, which could explain why participants interpreted LRs more accurately than RMPs. In other words, the present work could have overlooked the extent to which the social context influenced how police interpreted our experimental manipulations. Psychological experiments do not exist within a social vacuum, as even a survey experiment involves an interaction between an experimenter and a participant (Hilton, 1995; Hilton & Slugoski, 2001). People within any social context are attempting to figure out what the other person is saying, and this assumption applies to an experiment where a researcher is communicating information to a participant.

Specifically, assumptions of conversational norms explained previous inconsistencies in participants' responses to the same information communicated in different ways (Hilton, 1995; Hilton & Slugoski, 2001). In our study, LRs could encourage police to consider aspects that increase the likelihood the suspect is guilty, whereas RMPs could encourage police to consider aspects that another person could have committed the crime. LRs might increase the likelihood of processing the evidence as incriminating based on the language "100 times more likely," whereas the RMP condition might have enhanced the processing the evidence as exculpatory based on the language "one in 100." Instead of a mere statistical reframing, this format change could be encouraging participants to think differently about guilt or innocence, thereby rendering different patterns of guilt judgments due to the social context of the experimental paradigm. Thus, police could perceive LRs as conveying qualitatively different information than RMPs due to their motivation to cooperate with the experimenter. Future research could manipulate statistical formats of evidence strength

and an explicit prompt to consider guilt or innocence to further parse out how the social context of conversational norms is affecting participants' guilt judgments. Despite limitations, this study provided an important step in replicating and extending previous psychological research regarding how people interpret statistical formats of evidence from jury decision-making to a police investigation context.

Conclusion

Inaccurate evaluations of evidence can lead to wrongful convictions, so there is a need for police investigators to accurately interpret evidence. This study examined the effects of evidence strength format and evidence type on police officers' guilt judgments, extending the literature on interpretations of statistical formats from jury research to the context of an investigation. Overall, our findings helped clarify that LRs are the best format for statistically conveying evidence strength to police officers.

There is a surprising lack of research examining how police investigators interpret LRs, especially considering forensic experts use LRs to communicate evidence strength (Association of Forensic Science Providers, 2009) and the numerous studies that define and calculate the LRs associated with different eyewitness lineup procedures (e.g., Steblay et al., 2011; Wells & Lindsay, 1980). Police officers, in general, were relatively accurate at interpreting evidence strength, except for when evidence strength is presented in an RMP format for DNA evidence. Thus, forensic experts and researchers could benefit from conveying evidence strength in an LR format to police whenever possible, as well as police could benefit from conveying evidence strength in an LR format to other police officers when conducting an investigation.

References

Ask, K., Rebelius, A., & Granhag, P. A. (2008). The 'elasticity' of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology*, *22*(9), 1245–1259. https://doi.org/10.1002/acp.1432

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*, 161–164. https://doi:10.1016/j.scijus.2009.07.004

Bornstein, B. H., Golding, J. M., Neuschatz, J., Kimbrough, C., Reed, K., Magyarics, C., & Luecht, K. (2017). Mock juror sampling issues in jury simulation research: A meta-analysis. *Law and Human Behavior*, *41*(1), 13–28. https://doi.org/10.1037/lhb0000223

Borstein, B. H., & McCabe, S. G. (2004). Jurors of the absurd? The role of consequentiality in jury simulation research. *Florida State University Law Review*, *32*(2), 443–468.

Devine, D. J., & Macken, S. (2016). Scientific evidence and juror decision making: Theory, empirical research, and future directions. In B. H. Bornstein & M. K. Miller (Eds.), *Advances in Psychology and Law: Volume 2* (pp. 95–139). Springer International Publishing. https://doi.org/10.1007/978-3-319-43083-6_4

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *The American Journal of Trial Advocacy*, *16*, 361.

Hilton, D. J. (n.d.). *The social context of reasoning: Conversational inference and rational judgment*. 24.

Hilton, D. J., & Slugoski, B. R. (2001). Conversational processes in reasoning and explanation. In A. Tesser & N. Schwarz (Eds.), *Blackwell Handbook of Social Psychology: Intraindividual Processes* (pp. 181–206). Blackwell Publishers Inc. https://doi.org/10.1002/9780470998519.ch9

Horgan, A. J., Russano, M. B., Meissner, C. A., & Evans, J. R. (2012). Minimization and maximization techniques: Assessing the perceived consequences of confessing and confession diagnosticity. *Psychology, Crime & Law*, *18*(1), 65–78. https://doi.org/10.1080/1068316X.2011.561801

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, *240*, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human*

Behavior, *37*(3), 197–207. https://doi.org/10.1037/lhb0000027

Russano, M. B., Meissner, C. A., Narchet, F. M., & Kassin, S. M. (n.d.). *Investigating true and false confessions within a novel experimental paradigm*. *16*(6), 6.

Scherr, K. C., Redlich, A. D., & Kassin, S. M. (2020). Cumulative disadvantage: A psychological framework for understanding how innocence can lead to confession, wrongful conviction, and beyond. *Perspectives on Psychological Science*, *15*(2), 353–383. https://doi.org/10.1177/1745691619896608

Scudder, N., Kelty, S. F., Busby Grant, J., Montgomerie, C., Walsh, S. J., Robertson, J., & McNevin, D. (2020). *Differing perception of DNA Evidence and intelligence capabilities in criminal investigations* [Preprint]. LIFE SCIENCES. https://doi.org/10.20944/preprints202002.0004.v1

Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors' use of probabilistic evidence. *Law and Human Behavior*, *20*, 49–82. http://dx.doi.org/10.1007/BF01499132

Sommers, S. R., & Douglass, A. B. (2007). Context matters: Alibi strength varies according to evaluator perspective. *Legal and Criminological Psychology*, *12*(1), 41–54. https://doi.org/10.1348/135532506X114301

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*(1), 99–139. https://doi.org/10.1037/a0021650

The National Registry of Exonerations (2022). Retrieved, https://www.law.umich.edu/

Thompson, W. C., Grady, R. H., Lai, E., & Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law,*

Probability and Risk, *17*(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Thompson, W. C., Kaasa, S. O., & Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence. *Journal of Empirical Legal Studies*, *10*(2), 359-398.

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*(2), 45–75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Appendix

Evidence Manipulations

Language for these manipulations was adapted from Martire and colleagues (2014), Thompson and Newman (2015), and Wells and Lindsay (1980).

DNA Evidence

1. Neutral evidence: After the initial stage of the investigation, new evidence emerged. Particularly, DNA evidence was found.
   The forensic expert determined it is inconclusive as to whether the DNA sample from the crime scene matched the suspect's DNA because the amount of DNA left at the crime scene was too small to compare.
2. LR: After the initial stage of the investigation, new evidence emerged. Particularly, DNA evidence was found.
   The forensic expert determined the characteristics of the DNA evidence are 100 times more likely if the DNA evidence originated from the suspect than if the evidence originated from a randomly chosen person.
3. RMP: After the initial stage of the investigation, new evidence emerged. Particularly, DNA evidence was found.
   The forensic expert concluded one in 100 people have a DNA profile that is consistent with the DNA evidence found at the crime scene. That means that there is one chance in 100 of finding a consistent profile in a randomly chosen person.

Fingerprint Evidence

1. Neutral evidence: After the initial stage of the investigation, new evidence emerged. Particularly, fingerprint evidence was found.
   The forensic expert determined it is inconclusive whether the fingerprint sample from the crime scene matches the suspect because the fingerprint left at the crime scene was too small to compare.
2. LR: After the initial stage of the investigation, new evidence emerged. Particularly, fingerprint evidence was found.
   The forensic expert determined the characteristics of the fingerprint evidence are 100 times more likely if the fingerprint evidence originated from the suspect than if the evidence originated from a randomly chosen person.
3. RMP: After the initial stage of the investigation, new evidence emerged. Particularly, fingerprint evidence was found.
   The forensic expert concluded one in 100 people would have a fingerprint that is consistent with the fingerprint sample found at the crime scene. That means that there is one chance in 100 of finding a consistent fingerprint pattern from a randomly chosen person.

Eyewitness Evidence

1. Neutral: After the initial stage of the investigation, new evidence emerged. Particularly, an eyewitness identified the suspect as the perpetrator.
   The investigative expert determined it is inconclusive whether the suspect in the

lineup matched the culprit because the eyewitness did not have a good view of the suspect.

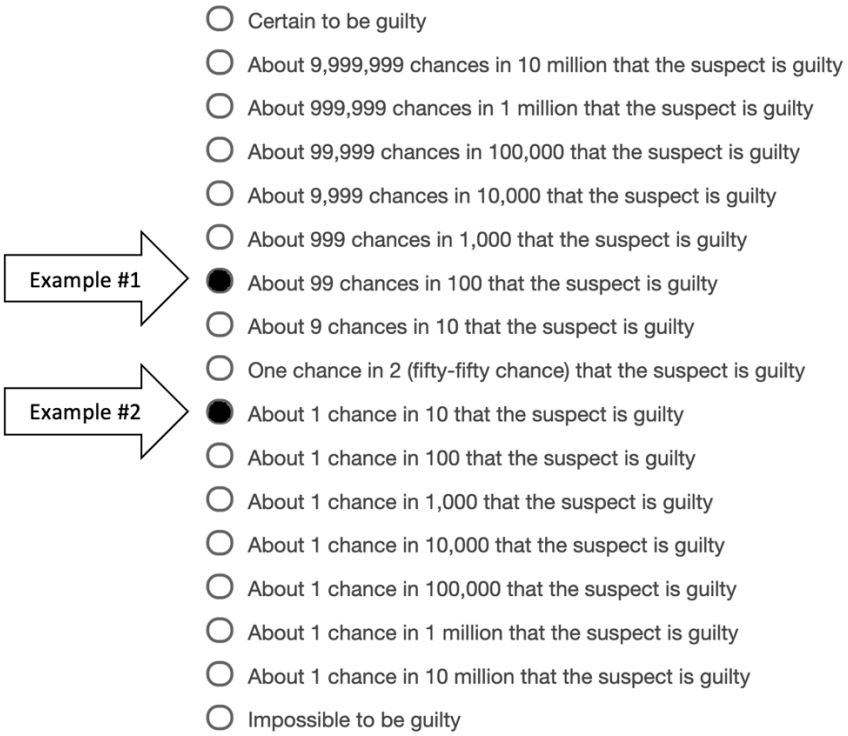
2. LR: After the initial stage of the investigation, new evidence emerged. Particularly, an eyewitness identified the suspect as the perpetrator. The investigative expert concluded the eyewitness ID evidence is 100 times more likely to have occurred if the suspect was at the crime scene than if the suspect was a randomly chosen person.
3. RMP: After the initial stage of the investigation, new evidence emerged. Particularly, an eyewitness identified the suspect as the perpetrator. The investigative expert concluded one in 100 people would be chosen from the lineup by the eyewitness. That means there is a one chance in 100 of this eyewitness identifying the suspect if the suspect was a randomly chosen person.

Instructions for Interpreting the Log-Scale

On the next page, you will be asked to judge the chances that the suspect is guilty of a crime based on the information you just read. You will use a scale like the one below to tell us your opinion.

For example, if you think there are 99 chances in 100 that the suspect is guilty (and only one chance in 100 that he is innocent)-then you should mark the box next to the statement "About 99 chances in 100 that the suspect is guilty" (see Example #1 below).

On the other hand, if you think there is only 1 chance in 10 that the suspect is guilty (and 9 chances in 10 that he is innocent), then you should mark the box next to the statement "About 1 chance in 10 that the suspect is guilty" (see Example #2 below).



This scale allows you to express a wide range of opinions about the chances of

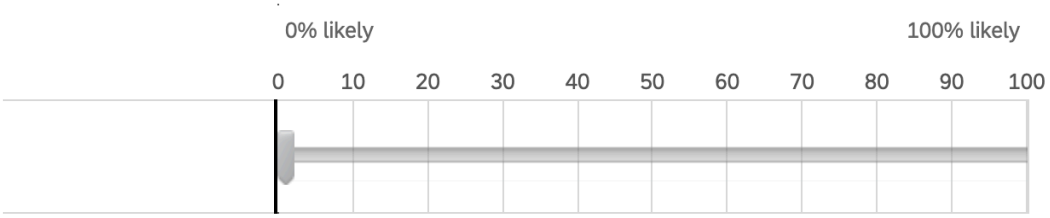
guilt, from extremely high to extremely low.

Use extremely high values on the scale, such as 999,999 chances in 1 million or higher, if you believe it is nearly certain that the suspect is guilty. Use extremely low values on the scale, such as 1 chance in 1 million or lower, if you believe it is nearly impossible that the suspect is guilty. Mark whatever box is closest to your belief about the chances the suspect is guilty.

If you believe that a suspect is certain to be guilty (in other words, there is no chance he could be innocent) you can indicate that by marking the top box on the scale, which says "Certain to be guilty." If you believe that it is impossible that a suspect is guilty (in other words, there is no chance he could be guilty), you can mark the bottom box on the scale, labeled "Impossible that he is guilty."

Percentage Scale Guilt Judgment

Considering the information in the case, what is the likelihood the suspect is guilty of committing murder from 0 (0% likely) to 100 (100% likely)?



Suspicion, Attention, and Manipulation Check

1. Please indicate what you knew about this study before participating. [open response]
2. Do you think the purpose of this study is obvious?
    a. Yes
    b. No
3. Please indicate what research questions you believe might be under investigation in this study.

--Page Break--

1. Did you pay attention to this study?
2. Did you read the full case scenario?

--Page Break--

1. What type of evidence did you read about?
    a. DNA
    b. Fingerprint
    c. Eyewitness
    d. I do not remember.

1a. [If participant choses DNA] What was the format of the evidence you read about?

a. That the evidence was inconclusive
b. That the characteristics of the DNA evidence are 100 times more likely if the DNA evidence originated from the suspect than if the evidence

originated from a randomly chosen person
c. That one in 100 people have a DNA profile that is consistent with the DNA evidence found at the crime scene
d. I do not remember

1b. [If participant choses fingerprint] What was the format of the evidence you read about?

a. That the evidence was inconclusive
b. That the characteristics of the fingerprint evidence are 100 times more likely if the fingerprint evidence originated from the suspect than if the evidence originated from a randomly chosen person
c. That one in 100 people would have a fingerprint that is consistent with the fingerprint sample found at the crime scene
d. I do not remember

1c. [If participant choses eyewitness] What was the format of the evidence you read about?

a. That the evidence was inconclusive
b. That the eyewitness ID evidence is 100 times more likely to have occurred if the suspect was at the crime scene than if the suspect was a randomly chosen person
c. That one in 100 people would be chosen from the lineup by the eyewitness
d. I do not remember

Demographic Questionnaires

Screener Question

Which of the following categories best describes the industry you primarily work in?

- Agriculture, Forestry, Fishing and Hunting
- Utilities
- Computer and Electronics Manufacturing
- Transportation and Warehousing
- Software
- Real Estate, Rental and Leasing
- Health Care and Social Assistance
- Primary/Secondary (K-12) Education
- Hotel and Food Services
- Legal Services
- Law Enforcement
- Retail
- Finance and Insurance
- College, University, and Adult Education
- Government and Public Administration
- Military
- Retired
- Unemployed
- Student

- Other

End of Survey Demographic Questions

1. How many years of experience as a police officer do you have? [open response]
2. What type of law enforcement agency do you currently work for?
    a. Local
    b. State
    c. Federal
    d. Other (please describe)
3. What is your current ranking in your department?
    a. Chief of Police
    b. Deputy Chief
    c. Detective/investigator
    d. Patrol Officer
    e. Lieutenant
    f. Police Officer
    g. Sheriff
    h. Sergeant
    i. Captain
    j. Other (please describe)
4. What is your gender?
    a. Female
    b. Male
    c. Other
5. What is your age? ___
6. Please indicate your ethnicity/race:
    a. Black or African American
    b. Asian
    c. White or Caucasian
    d. Latina/o
    e. Native American
    f. Indian
    g. Multi-ethnic (Please indicate your ethnicity/race.) ____
    h. Other (Please indicate your ethnicity/race.) ____
7. What is your highest level of education?
    a. Some high school, no diploma
    b. High school graduate, diploma or the equivalent (for example: GED)
    c. Some college credit, no degree
    d. Trade/technical/vocational training
    e. Associate degree
    f. Bachelor's degree
    g. Master's degree
    h. Professional degree
    i. Doctorate degree
8. Generally speaking, would you describe your political views as…

   a. Very Liberal
   b. Somewhat Liberal
   c. Moderate
   d. Somewhat Conservative
   e. Very Conservative
9. Generally speaking, would you describe your political party as…
   a. Republican
   b. Democrat
   c. Independent
   d. Other ____
10. Do you have any questions or comments for the research team?

Chapter 4: Paper #3

Decreasing Biased Guilt Judgments During an Investigation with Social Norms

Jean J. Cabell[1] and Yueran Yang[2]

[1]Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno

[2]Psychology Department and Interdisciplinary Social Psychology Ph.D. Program, University of Nevada, Reno.

Author Note

Jean J. Cabell, https://orcid.org/0000-0002-2362-0419

Yueran Yang, https://orcid.org/0000-0001-9261-5608

We have no known conflicts of interest to disclose.

This work was supported by the Russell J. and Dorothy S. Bilinski Fellowship Fund, a program of the Bilinski Educational Foundation, and a Research and Materials Grant awarded by the Graduate Student Association of the University of Nevada, Reno.

Correspondence concerning this article should be addressed to Jean J. Cabell, Interdisciplinary Social Psychology Ph.D. Program, Mailstop 1300, University of Nevada, Reno, 1664 N. Virginia St., Reno, NV 89557. Email: jcabell@nevada.unr.edu

Abstract

Objective: Police investigators evaluate multiple pieces of evidence when forming their guilt judgments, but errors in evaluating evidence to judge suspect guilt could lead to wrongfully incriminating an innocent suspect. The order police encounter evidence is one factor that could bias police judgments of suspect guilt. Ideally, police judgments of suspect guilt should follow the Bayesian Cognitive Model, such that the order police encounter evidence should not affect their final guilt judgments. However, there is mixed evidence as to whether police over-weigh the first piece of evidence (confirmation bias) or the last piece of evidence (recency bias) in their guilt judgments. Social norms that promote thorough investigations could minimize bias if it occurs when compared to social norms that promote efficient investigations. **Hypotheses:** We offered competing hypotheses for the effects of evidence order on evaluations of evidence and judgments of suspect guilt. Evidence order should not affect guilt judgments according to the Bayesian Cognitive Model. Conversely, police could over-weigh the initial piece of evidence (confirmation bias) or the last piece of evidence (recency bias) in their judgments of suspect guilt. **Method:** We recruited a police sample (Study 1) and an Amazon Mechanical Turk (MTurk) sample (Study 2) and randomly assigned them to a 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (social norms: efficiency vs. thoroughness) between-subjects factorial design. **Results:** Police participants and MTurk participants displayed a recency bias in their guilt judgments, but this bias was minimized by social norms of thoroughness. **Conclusion:** Police departments should aim to increase social norms of thoroughness to minimize bias when using evidence to judge suspect guilt.

Decreasing Biased Guilt Judgments During an Investigation with Social Norms

On June 13th, 1996, an 18-year-old woman named Angie Dodge was raped and murdered. Police investigators identified Christopher Tapp as a potential suspect and coerced him to confess to the crimes (Murphy, 2019). DNA evidence collected from the crime scene did not match Tapp, but police investigators ignored this DNA evidence and maintained that Tapp was guilty. Tapp was convicted. It was not until 2019 that analysts working with the Idaho Innocence Project identified DNA from another man, Brian Dripps, that matched the DNA sample from the crime scene (Otterbourg, 2021). Tapp was proven to be innocent of committing these crimes and was exonerated after over 20 years in prison for a crime he did not commit. This case is an example of police failure in accurately evaluating evidence when judging a suspect's guilt and how this inaccuracy can lead to the conviction and imprisonment of an innocent person.

Police evaluate evidence to judge the likelihood of a suspect's guilt, perhaps even multiple times, to make decisions throughout an investigation (Ask & Alison, 2010). For example, police might base their decisions on their judgments of suspect guilt, such as whether to gather more evidence, interrogate a suspect, arrest a suspect, or refer a case to a prosecutor. However, police investigations could go awry if innocent people, such as Christopher Tapp, are misidentified as suspects because police's erroneous judgments could lead them to make decisions that end in wrongfully convicting innocent people (Scherr et al., 2020). Thus, evaluating multiple pieces of evidence properly is a key factor in enabling police investigators to accurately judge a suspect's guilt.

The order in which police encounter evidence is one factor that could affect police judgments of suspect guilt. Ideally, police evaluations of evidence and judgments of

suspect guilt would follow the Bayesian Cognitive Model because this model assumes accuracy and rationality (Druckman & McGrath, 2019). This model of cognition is based on Bayes' theorem, posterior $\propto$ prior $\times$ new data, which states the optimal method to combine new information with old information is to simply combine prior information with new information to form posterior beliefs (Druckman & McGrath, 2019; Edwards, 1962). This model assumes new information is evaluated independently of prior information, and therefore the order police encounter information should not affect their posterior beliefs (i.e., final judgments of suspect guilt).

However, evidence order can bias evidence evaluations and guilt judgments (e.g., Charman et al., 2016; Price & Dahl, 2013). There are mixed findings on whether people are biased by the first piece of information they encounter (i.e., a confirmation bias effect; Charman et al., 2016) or by the last piece of information they encounter (i.e., a recency bias effect; Price & Dahl, 2013). Further, little research has examined potential methods to minimize the distorting influence of evidence order. One factor that could potentially reduce this bias is social norms because norms prioritizing a thorough investigation led to greater processing of later evidence compared to social norms of an efficient investigation (Ask et al., 2011). Therefore, this research attempts to address the following overarching research questions: 1) How does evidence order affect evaluations of evidence and suspect guilt? 2) How do social norms affect evaluations of evidence and suspect guilt?

Evidence Order

During criminal investigations, police generally need to evaluate multiple pieces of evidence (Ask & Alison, 2010). This section reviews three primary theories of

information processing that can offer predictions for the effects of evidence order on police evaluations of evidence and judgments of suspect guilt during an investigation. We begin by reviewing the Bayesian Cognitive Model as a normative model that predicts how people *ought* to make decisions. Then, we review confirmation bias and recency bias as descriptive models that can offer predictions as to how people *actually* make judgments and decisions.

Bayesian Cognitive Model

The Bayesian Cognitive Model can predict how people update their beliefs when making decisions under uncertainty (Chater et al., 2006; Druckman & McGrath, 2019; Edwards, 1962). This model of cognition is based on Bayes' theorem: posterior $\propto$ prior $\times$ new data. Bayes' theorem specifies the optimal method to combine new information with old information, such that the posterior probability of an outcome is proportional to the prior probability of the outcome multiplied by the likelihood of the outcome (Bowers & Davis, 2012; Griffiths et al., 2008). Applying this theorem to cognition, people form their posterior beliefs by combining their prior expectations with new information (Edwards, 1962). Bayesian models typically assume that people are motivated by accuracy to arrive at the correct conclusion (Druckman & McGrath, 2019) and that they are rational when updating their beliefs (i.e., they combine all relevant information effectively; Charman et al., 2009; Jacobs & Kruschke, 2011). Therefore, evaluations of new information are independent of prior information and prior judgments (Blair & Rossmo, 2010; Druckman & McGrath, 2019). This model offers three relevant predictions.

First, this model predicts the order police encounter evidence should not affect

their final guilt judgments. Bayesian models predict posterior (i.e., final) guilt judgments will be the same regardless of the order that police encounter evidence. For example, police's final guilt judgments should be the same regardless of whether they encounter ambiguous evidence first or incriminating evidence first because Bayesian models assume independence between information. Therefore, police should simply add up the evidence to render a final guilt judgment and the order they encounter the evidence should not affect their final guilt judgments.

Second, this model predicts final guilt judgments should be the same as initial guilt judgments if the new evidence is ambiguous. A key component of this updating process is that police evaluations of evidence should not be influenced by prior information and judgments due to the assumption that all information is evaluated independently. If a prior (i.e., initial) guilt judgment is strong and new evidence is ambiguous, a posterior (i.e., final) guilt judgment should be as strong as the initial guilt judgment. Importantly, the final guilt judgment should not be *stronger* than the prior, but rather it should be the *same* as the prior because ambiguous evidence does not offer any new information. Current literature either does not measure initial guilt beliefs (e.g., Kassin et al., 2003) or does not directly compare initial guilt beliefs with final guilt judgments (e.g., Charman et al., 2017), therefore it is unknown whether police investigators are using Bayesian processes when updating their final guilt judgments.

Finally, this model predicts that ambiguous evidence should be evaluated the same regardless of what order police encounter it because Bayesian models predict that new evidence is evaluated independently from prior beliefs. Police, therefore, should evaluate ambiguous evidence the same regardless of whether they encounter the evidence

before or after incriminating evidence.

Confirmation Bias

Confirmation bias refers to the biased tendency to search for, remember, or interpret new information consistently with prior beliefs (Nickerson, 1998). The latter tendency to interpret new information in a way that confirms previous beliefs is most relevant to this study. Confirmation bias suggests that the first evidence police uncover during an investigation could bias evaluations of later evidence in the direction of the first piece of evidence. For example, Charman and colleagues (2017) found that the initial piece of evidence police and student participants encountered predicted their final guilt judgments when they were presented with multiple pieces of evidence, suggesting a confirmation bias effect.

Confirmation bias assumes people desire to arrive at their preferred conclusion rather than the correct conclusion (Druckman & McGrath, 2019). This framework also assumes new information reinforces initial beliefs (Druckman & McGrath, 2019). Therefore, people are more likely to evaluate new information in a manner that confirms their existing beliefs, especially when the new information is ambiguous, which could reinforce and strengthen their emerging conclusions.

The assumption of a confirmation bias offers the following predictions. First, final guilt judgments will be higher when police encounter incriminating evidence first and ambiguous evidence second (vs. ambiguous evidence first and incriminating evidence second) because they should evaluate the new evidence in a manner that confirms their existing beliefs. Second, police should evaluate ambiguous evidence as neither incriminating nor exonerating when they do not have a strong initial guilt belief but they

should evaluate the ambiguous evidence in the direction of their initial beliefs (Charman et al., 2016), thus police should evaluate ambiguous evidence as less incriminating when they encounter the ambiguous evidence first and incriminating evidence second (vs. incriminating evidence first and ambiguous evidence second). Third, judging ambiguous evidence as incriminating could then strengthen police's beliefs that the suspect is guilty, leading to a final guilt judgment that is stronger than their initial guilt judgment due to confirmation bias (i.e., a bias snowball effect; Charman et al., 2017).

Recency Bias

In contrast to confirmation bias, a recency bias suggests the last piece of information has a greater effect on a final judgment than other pieces of information (Carlson & Russo, 2001; Dahl et al., 2009). For example, Charman and colleagues (2016) had participants read a trial case summary, evaluate two pieces of evidence, and render a final guilt judgment. The authors manipulated the order of the evidence (DNA before vs. after an ambiguous alibi) and the valence of the DNA evidence (incriminating vs. exonerating). Participants' final guilt judgments were most consistent with the last piece of evidence they evaluated, suggesting a recency effect (Charman et al., 2016).

A recency bias suggests the following predictions. First, police who display a recency bias should have final guilt judgments that are driven by the last piece of evidence such that they should have higher guilt judgments when encountering ambiguous evidence first and incriminating evidence second (vs. encountering incriminating evidence first and ambiguous evidence second). Second, police's final guilt judgments should be lower than their initial guilt judgments when they encounter incriminating evidence before ambiguous evidence because the most recent piece of

information should be driving their final guilt judgments, despite ambiguous evidence not adding any new information. Because the recency bias only affects final judgments, there are no predicted effects of recency bias on evaluations of ambiguous evidence.

Contradictory Findings of Evidence Order on Guilt Judgments

Past literature shows mixed findings on the effects of order on guilt judgments; thus, it is unclear whether police engage in confirmation bias or recency bias when judging suspect guilt based on multiple pieces of evidence. One possible reason for these mixed findings could be that researchers showed participants one piece of evidence typically that is typically incriminating (e.g., eyewitness ID evidence) and one type of evidence that is typically exculpatory (e.g., alibi; Charman et al., 2016; Dahl et al., 2009; Price & Dahl, 2013). No confirmation bias effects were found in these studies that used one piece of evidence that is typically incriminating and one piece of evidence that is typically exonerating. In a later study conducted by Charman and colleagues (2017), initial beliefs of guilt did predict evaluations of potentially incriminating evidence (i.e., composites, handwriting, and informant evidence), but not of potentially exculpatory evidence (i.e., alibi evidence). Thus, confirmation bias might only be found when all types of evidence are potentially incriminating. However, Charman and colleagues (2017) did not vary evidence order or test whether final guilt judgments were different than initial guilt judgments, so it is unknown whether investigators are prone to confirmation bias when they encounter ambiguous and incriminating evidence or the extent to which these effects occur.

Social Norms as a Moderator of Order Effects

Social influence is one potential method to decrease biases that result from order

effects (i.e., confirmation or recency bias). The most prominent and instrumental form of social influence is conforming to relevant group norms (Hogg, 2010). Conformity occurs when people look to the behaviors of others and match their behavior to others (Cialdini & Goldstein, 2004). People rely on the behaviors of others to form an accurate interpretation of reality, to gain information, or to obtain approval from others (Cialdini & Goldstein, 2004; Risinger et al., 2002). Group norms could influence police because people tend to conform to others and adopt similar goals as those endorsed by significant others (Cialdini & Goldstein, 2004; Shah, 2003).

Social norms prioritizing a thorough (vs. efficient) investigation led to increased consideration of later evidence, whereas norms prioritizing an efficient investigation led investigators to discount contradictory evidence discovered later in an investigation (Ask et al., 2011). Thus, investigators in the thoroughness norms condition judged a suspect's guilt in the direction of the evidence discovered later in the case, suggesting they processed the later evidence more in the thoroughness norms condition than in the efficient condition. However, Ask and colleagues (2011) only manipulated incriminating and exculpatory evidence (i.e., contradictory evidence). Because biases are more likely to occur when there is ambiguous evidence (Risinger et al., 2002; Snook, 2000), it is still unknown whether social norms could minimize bias from order effects when there is ambiguous evidence. Ask and colleagues (2011) also did not manipulate the order of evidence or measure how guilt beliefs changed after each piece of evidence, so it is unknown how police updated their beliefs. We expect participants exposed to norms of thoroughness (vs. efficiency) should be less biased in their evidence evaluations and guilt judgments, and therefore less influenced by the order of evidence.

Hypotheses

We predict three competing hypotheses according to the Bayesian Cognitive Model, confirmation bias, and recency bias (see Table 1). We also predicted a moderating effect of social norms, such that social norms of thoroughness (vs. efficiency) should minimize the biased effects predicted by confirmation bias or recency bias if these biased effects occur. We executed two studies to test these hypotheses. Study 1 recruited police but suffered from a small sample size. Study 2 recruited a larger sample size from laypeople on Amazon Mechanical Turk (MTurk) to account for the small sample size and to offer a comparison as to whether MTurk samples can generalize to police officers. Each study is discussed in turn.

Table 1

Competing hypothesized order effects

Competing Hypotheses	**Bayesian Cognitive Model**	**Confirmation Bias**	**Recency Bias**
H1: Final Guilt Judgments	Mean final guilt judgments will be the same between order conditions	Final guilt judgments will be **higher** in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*)	Final guilt judgments will be **lower** in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*)
H2: Updating Guilt Judgments	Mean initial guilt judgments will not differ from mean final guilt judgments in the *incriminating first-ambiguous second* condition	Mean initial guilt judgments will be **lower** than mean final guilt judgments in the *incriminating first-ambiguous second* condition	Mean initial guilt judgments will be **higher** than mean final guilt judgments in the *incriminating first-ambiguous second* condition
H3: Evaluations of Ambiguous Evidence	Mean evaluations of ambiguous evidence will be the same between order conditions	Mean evaluations of ambiguous evidence will be **guiltier** in the *incriminating first-ambiguous second* condition (vs. *ambiguous first-incriminating second*)	N/A

Study 1

The first study is an online experiment examining the effects of evidence order and social norms on evaluations of evidence and judgments of suspect guilt using a police sample. We randomly assigned participants to a 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (social norms: efficiency vs. thoroughness) x 2 (type: DNA incriminating-eyewitness ambiguous vs. eyewitness incriminating-DNA ambiguous) between-groups factorial design. The evidence order factor manipulated whether participants read about the strongly

incriminating evidence before or after the ambiguous evidence. The norms factor manipulated whether participants read that their peers endorsed norms of investigative efficiency or norms of investigative thoroughness.

The type factor manipulated whether participants read about incriminating DNA evidence and ambiguous eyewitness ID evidence or incriminating eyewitness ID evidence and ambiguous DNA evidence. We chose DNA evidence as one type of evidence because past research used incriminating DNA evidence (Charman et al., 2016). We wanted another piece of evidence that is typically used to incriminate suspects, so we chose eyewitness ID evidence as the second piece of evidence. Although there are no research questions related to evidence type, we included this factor to account for the potential confound between evidence type and evidence strength. For example, if our design was such that the incriminating evidence was always DNA and the ambiguous evidence was always eyewitness, we would not be able to parse out whether any order effects were related to the type of evidence or the strength of evidence. Based on the Bayesian Cognitive Model, confirmation bias, and recency bias, we were interested in the effects of evidence order when one piece of evidence is incriminating and one piece of evidence is ambiguous. Thus, the study employed a 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (social norms: efficiency vs. thoroughness) design and counterbalanced evidence type across order conditions to account for the potential confound between evidence strength and evidence type within the evidence order factor.

Study 1 Method

Participants

We recruited 75 U.S. sworn police officers using a hybrid-snowball technique. We contacted police officers within our networks and requested they distribute the survey within their departments and to other officers who might be interested. To recruit additional officers outside of the snowball sample, we registered the study with the Professional Research Pool for Criminal Justice Science (n.d.), contacted police chiefs and personnel from a random sample of 200 departments from a list of 15,810 U.S. law enforcement agencies (LEAR, 2017)[2], and posted the study on Police1, which is a news website with a police audience (police1.com). We recruited police between September 2021 through December 2021.

Our final sample was 62 participants after removing participants who agreed to participate but did not answer any items and participants who did not complete the suspicion check, manipulation check, and demographic questions. The sample self-identified their gender as 83% men, 10% women, and 7% other. Participants self-identified their race and ethnicity as 78% Caucasian, 5% African American, 3% Latino/a, 3% Asian, 2% Native American, and 10% multi-ethnic. Participants' mean age was 42.97 (*SD* = 10.84; range = 21-66). On average, participants had 17.8 years of law enforcement experience and self-identified their department as 86% local, 8% state, and 5% other. Participants' self-reported ranking within their departments was 13% Chief of Police, 13% Detective/Investigator, 13% patrol officer, 3% Lieutenant, 20% police officer, 20%

[2] Some randomly selected departments creased operate, or they did not list contact information. Of these 200 randomly selected departments, 160 were successfully contacted (i.e., had a working email or phone number) and 4 departments agreed to participate.

Sergeant, 3% Captain, and 13% other.

Materials

Norms Manipulation. Participants read a cover story that indicated a large majority (i.e., more than 80%) of those who previously took the survey agreed with a list of six statements regarding the quality of a good police investigator in previous research (Ask et al., 2011). Two statements were identical between conditions: "A good investigator knows how to make good use of his/her prior practical experience" and "A good investigator has good communication skills." The remaining four statements varied by *efficiency* or *thoroughness* conditions. In the *efficiency norms* condition, participants read statements related to investigative *efficiency* (e.g., "A good investigator often sees a solution to a crime early in the investigation"). In the *thoroughness norms* condition, participants will read statements related to investigative *thoroughness* (e.g., "A good investigator should avoid premature conclusions about a crime"). Participants rated how much they agreed with these statements from 1 (*strongly disagree*) to 9 (*strongly agree*). The purpose of these measures was to ensure participants read and encode the content of each statement but they are not part of the main analyses (see Appendix for the full list of statements).

Case Scenario. We adapted a written murder scenario from Charman and colleagues (2016), which was based on a real murder. A murder case was chosen because it is a crime that is feasible to have DNA and eyewitness ID evidence. The case scenario was pilot tested to ensure there were no ceiling or floor effects of the case scenario on guilt judgments (see Appendix for materials).

Incriminating Evidence. Participants in the *DNA incriminating-eyewitness*

ambiguous condition read a written summary that the DNA evidence strongly implicated guilt (see Appendix for materials). Participants in the *eyewitness incriminating-DNA ambiguous* condition read a written summary that the eyewitness ID evidence strongly implicated guilt. These conditions were pilot tested to ensure the evidence was strongly incriminating and that there was no significant difference in perceptions of how incriminating the evidence was by evidence type.

Ambiguous Evidence. Participants in the *DNA incriminating-eyewitness ambiguous* condition read a written summary that the eyewitness was not sure whether the suspect was in the lineup. Participants in the *eyewitness incriminating-DNA ambiguous* condition read a written summary that a forensic expert determined it was inconclusive as to whether the DNA sample from the crime scene matched the suspect. These conditions were pilot tested to ensure they were interpreted as neither incriminating nor exonerating and that there was no significant difference in perceptions of how incriminating the evidence was by evidence type.

Evidence Evaluations. Participants rated the extent to which each piece of evidence implied the suspect's guilt or innocence by answering "*To what extent does the _____ evidence imply Samuel Scott's innocence or guilt*" from 1 (*strongly implies innocence*) to 5 (*implies neither innocence nor guilt*) to 9 (*strongly implies guilt*). The type of evidence was in the blank. We asked this question after both the incriminating and ambiguous evidence to avoid suspicion, but our research question only concerns evaluations of ambiguous evidence.

Guilt Judgments. Participants rated the suspect's guilt twice by answering "*To what extent do you believe Samuel Scott is innocent or guilty*" from 1 (*completely*

innocent) to 5 (*neither guilty nor innocent*) to 9 (*completely guilty*) after both pieces of evidence to measure participants' initial and final guilt judgments.

Attention Check. Participants were asked whether they paid attention to both studies and whether they read the full case scenario.

Suspicion and Manipulation Check. We asked participants what they knew about the studies, if they had any suspicions, and to describe any suspicions (see Appendix). Next, we asked participants which series of statements they remembered from "Study 1" and showed them both sets of statements. Finally, we asked participants about the order they encountered the evidence and whether they read about incriminating DNA and ambiguous eyewitness ID evidence or incriminating eyewitness ID evidence and ambiguous DNA evidence in "Study 2."

Demographic Questionnaire. Participants answered questions regarding their number of years of law enforcement experience, the type of agency they currently work for (i.e., local, state, or federal), and their current department ranking. Participants also answered a basic demographic questionnaire assessing age, gender, race and ethnicity, education, and political identity. See Appendix for items.

Procedure

After reading and agreeing to a consent sheet, participants read a prompt that police officers were difficult to recruit so we were asking them to complete two ostensibly separate studies. However, this prompt was a cover story designed to decrease suspicion that the norms statements were related to their evidence evaluation and guilt judgment tasks (Ask et al., 2011). During the "first" study, we assigned participants to complete either the *efficiency* or the *thoroughness* norms manipulation measures. In the

"second" study, participants read the case scenario depicting a man named Samuel who was accused of murdering a young girl. Next, we randomly assigned participants to one of four groups: (1) *Incriminating DNA evidence first-ambiguous eyewitness evidence second*, (2) *Ambiguous eyewitness evidence first-Incriminating DNA evidence second, (3) Incriminating eyewitness evidence first-ambiguous DNA evidence second, and (4) Ambiguous DNA evidence first-incriminating eyewitness evidence second*. Participants read about the first piece of evidence and answered the evidence evaluation measure and the guilt judgment measure. Participants then read about the second piece of evidence, again answering the evidence evaluation measure and the guilt judgment measure. Finally, they answered the manipulation, attention, suspicion check, and demographics questionnaire. After completing the study, participants had the option to enter their email addresses on a separate survey to enter a raffle to win one of 45 $20 Amazon gift cards.

Results

This section reports results to test our hypothesized effects. All analyses were conducted using R.

Attention, Suspicion, and Manipulation Checks

Eight participants reported they either did not read or pay attention to the study. No participants correctly inferred the purpose of the norms manipulation or that the order of evidence varied. The manipulation checks revealed that 6 participants incorrectly remembered the norms manipulation, 11 participants incorrectly remembered the order they encountered the evidence, and 8 participants incorrectly remembered which type of evidence was incriminating. Some participants failed multiple attention or manipulation checks, so these numbers are not exclusive. We performed the analyses with ($N = 62$) and

without ($N = 42$) the data from the participants who reported they did not pay attention or who failed to correctly remember the manipulation. The pattern of results for the full and reduced sample was the same, except for the analyses on evaluations of ambiguous evidence.

Analyses conducted in this study were likely underpowered, as an a priori power analysis suggested recruiting 128 participants to detect a medium effect ($f = 0.25$) in a factorial ANOVA assuming a power of $1 - \beta = 0.80$ and a Type I error rate $\alpha = 0.05$. After calculating the post-hoc power based on the smallest effect size found in Study 1 ($\eta_P^2 = 0.06$), the study was underpowered for the full sample ($1 - \beta = 0.50$) and reduced sample ($1 - \beta = 0.36$) in the ANOVA analyses. However, the study did show sufficient post-hoc power in the full ($1 - \beta = 0.99$) and reduced sample ($1 - \beta = 0.96$) for the planned contrast analyses based on the smallest effect size ($d = 1.07$).

Preliminary Analyses

First, participants' agreement with the norms statements differed between the thoroughness and efficiency conditions $t(60) = 4.93$, $p < .001$. Participants were more likely to agree with the thoroughness statements ($M = 7.83$, $SD = 1.21$) than the efficiency statements ($M = 6.09$, $SD = 1.54$). However, these ratings did not significantly correlate with or qualify the main dependent variables and thus were not included in the analyses (Ask et al., 2011).

Second, we examined whether the effects of the norms manipulation or evidence order on the main dependent variables depended upon evidence type to determine whether these effects are confounded by evidence type in the full and reduced sample. Evidence type did not interact with evidence order or the norms manipulation on any of

our main dependent variables (*p*s > 0.21) and therefore was not included in the main analyses.

Final Guilt Judgments

We conducted a 2 (social norms: thorough vs. efficient) x 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) between-groups factorial ANOVA analysis on final guilt judgments to examine whether evidence order affected guilt judgments according to the Bayesian Cognitive Model, confirmation bias, or recency bias, as well as whether social norms decreased biased effects should we find any. Participants' final guilt judgments should not differ between order conditions according to the Bayesian Cognitive Model. However, participants' final guilt judgments could be higher (confirmation bias) or lower (recency bias) in the incriminating first-ambiguous second condition than in the ambiguous first-incriminating second condition.

There was a significant main effect of evidence order on final guilt judgments, $F(1, 57) = 5.17$, $p = .03$, $\eta_p^2 = 0.08$. Participants who encountered the ambiguous evidence first and incriminating evidence second had significantly higher final guilt judgments ($M = 6.49$, $SD = 1.26$) than participants who encountered incriminating evidence first and ambiguous evidence second ($M = 5.71$, $SD = 1.41$), which suggests a recency effect.

Although the two-way interaction did not approach conventional levels of significance, $F(1, 57) = 3.55$, $p = .065$, $\eta_p^2 = 0.06$, our planned contrasts revealed norms moderated the effect of evidence order. Specifically, participants' final guilt judgments were higher when they encountered ambiguous evidence first and incriminating evidence second ($M = 6.85$, $SD = 1.34$) than when they encountered incriminating evidence first

and ambiguous evidence second (M = 5.42, SD = 1.22), but only in the efficient norms condition, $t(57) = 2.99$, $p = .004$, $d = 1.08$, 95% CI [0.47, 2.38], and not in the thorough norms condition ($p = .79$). This interaction effect indicates norms of thoroughness eliminated police participants' recency bias.

After we removed participants who reported they did not pay attention or who failed to correctly remember the manipulation, there was only a marginally significant main effect of evidence order $F(1, 38) = 3.83$, $p = .06$, $\eta_p^2 = 0.09$. The effect remained in the same direction as reported above. There was still only a marginally significant interaction between the norms manipulation and evidence order, $F(1, 38) = 3.83$, $p = .057$, $\eta_p^2 = 0.09$. Planned contrasts revealed the interaction effect was in the same pattern as reported above.

Updating Guilt Judgments

We wanted to identify whether initial guilt judgments differed from final guilt judgments and had three competing hypothesized effects. According to the Bayesian Cognitive Model, participants' initial guilt judgments should not differ from their final guilt judgments in the incriminating first-ambiguous second condition because the ambiguous evidence offers no new information. Alternatively, police in the incriminating first-ambiguous second could render initial guilt judgments that are lower (confirmation bias) or higher (recency bias) than their final guilt judgments.

Participants only had one piece of evidence to consider when rendering their initial guilt judgments, and this piece of evidence differed in strength depending upon the evidence order condition. For example, participants in the ambiguous first-incriminating second condition provided their initial guilt judgments in response to the ambiguous

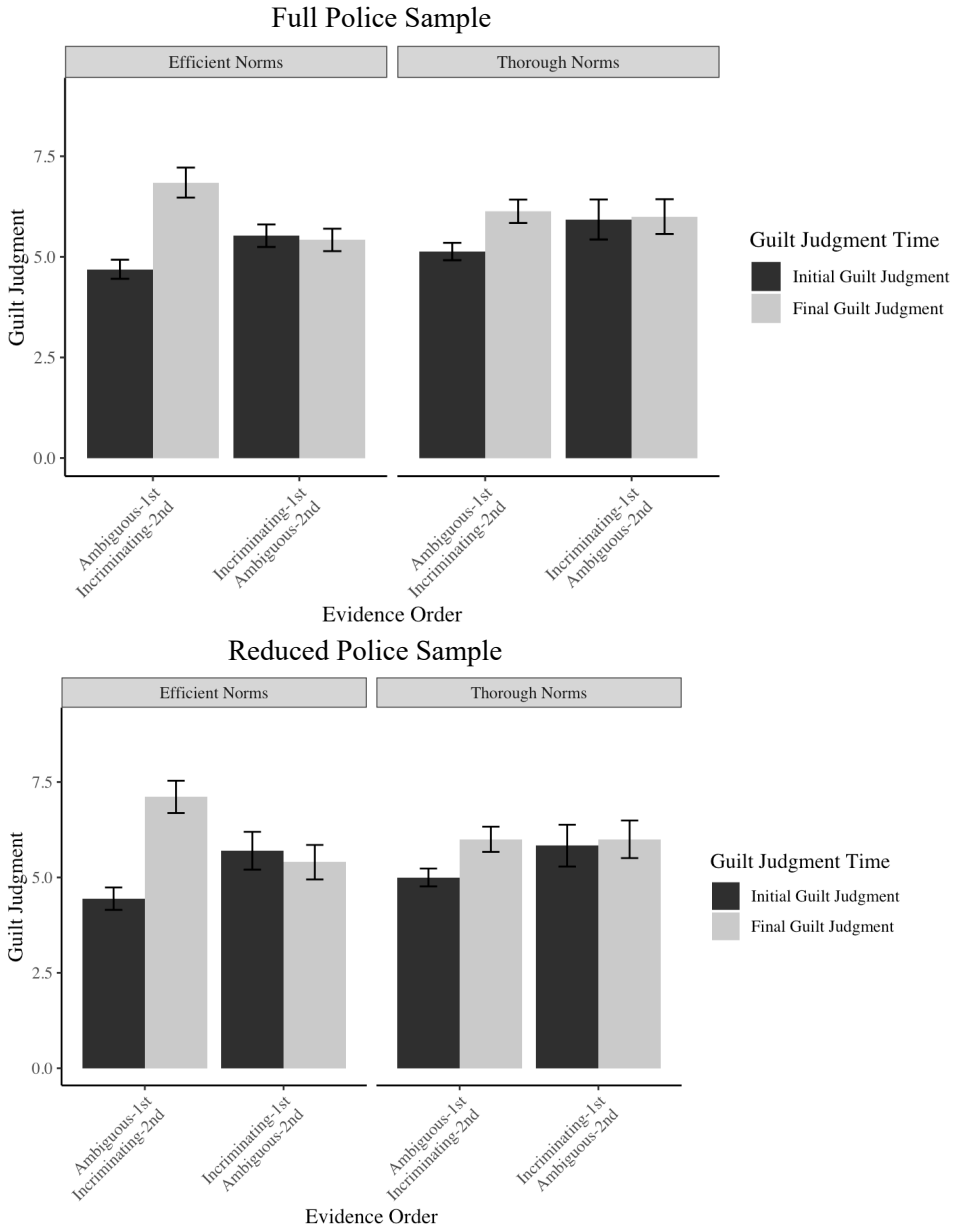
evidence. Participants had two (i.e., all) pieces of evidence to consider when rendering their final guilt judgments. Thus, participants providing their final guilt judgments had information from both pieces of evidence. We were interested in how participants integrated multiple pieces of evidence when updating their guilt judgments and used a linear mixed-effects model to examine how their guilt beliefs changed to incorporate the new information.

We used a linear mixed-effects model to examine whether initial and final guilt ratings differed by evidence order condition, and whether social norms moderated this effect. Fixed effects included evidence order (incriminating first-ambiguous second vs. ambiguous first-incriminating second), the norms manipulation (efficiency vs. thoroughness), guilt judgment time (i.e., initial guilt judgment vs. final guilt judgment), and their interactions. Random effects included the participants' ID.

The model revealed a significant main effect of guilt judgment time $F(1, 57) = 31.22$, $p < 0.001$ and an interaction between evidence order and guilt judgment time $F(1, 57) = 32.59$, $p < 0.001$, but these effects were qualified by a significant three-way interaction between evidence order, guilt judgment time, and social norms $F(1, 57) = 5.68$, $p = 0.02$ (see Figure 1).

Figure 1.

Effects of Evidence Order and Social Norms on Police Guilt Judgments



Note. The above figure depicts how the order police encountered evidence and social norms influenced how police updated their guilt judgments from the full and reduced samples. All initial guilt judgments were in response to the first piece of evidence, which differed in strength depending upon order condition. All final guilt judgments were after receiving both pieces of evidence. Participants' guilt judgments were on a 1-9 scale, such that answers above 5 suggest guilt, below 5 suggest innocence, and 5 suggest "*neither*

guilty nor innocent."

We initially hypothesized a potential interaction between social norms and guilt judgment time among participants in the incriminating first-ambiguous second condition if there was a recency bias or a confirmation bias, but instead found an interaction between social norms and guilt judgment time among participants in the ambiguous first-incriminating second condition. Contrasts revealed that when participants encountered ambiguous evidence before incriminating evidence, their initial guilt judgments (M = 4.69, SD = 0.85) were significantly lower than their final guilt judgments (M = 6.85, SD = 1.34) when primed with efficient norms, $t(57) = 7.20$, $p < .0001$, $d = 2.82$, 95% CI [1.55, 2.75]. This effect suggests participants who encountered ambiguous evidence first increased their guilt judgments in response to the incriminating evidence.

This effect was *weaker* when participants were primed with thorough norms such that their initial guilt judgments (M = 5.13, SD = 0.83) were significantly lower than their final guilt judgments (M = 6.13, SD = 1.13) when they encountered ambiguous evidence first and incriminating evidence second, $t(57) = 3.59$, $p < .001$, $d = 1.31$, 95% CI [0.44, 1.56]. Therefore, police participants were more conservative in updating their guilt judgments when they were primed with norms of thoroughness than norms of efficiency.

We found no significant interaction between social norms and guilt judgment time among participants in the incriminating first-ambiguous second condition. Specifically, there was no difference between initial and final guilt judgments, regardless of the norms manipulation (ps > 0.67), when participants encountered incriminating evidence before ambiguous evidence. This finding supports the Bayesian Cognitive Model because police

did not change their guilt judgments between time points, suggesting police interpreted the ambiguous evidence as offering no new information.

The pattern of results reported in the previous paragraph remained the same after we removed participants who reported they did not pay attention or who failed to correctly remember the manipulation, such that there was a significant main effect of guilt judgment type $F(1, 38) = 24.73$, $p < 0.001$ and interaction between evidence order and guilt judgment type $F(1, 38) = 18.74$, $p < 0.001$, but these effects were qualified by a significant interaction between evidence order, guilt judgment type, and social norms $F(1, 38) = 9.02$, $p = .003$ (see Figure 1). Follow-up contrasts indicated the pattern of the interaction effect was the same as reported in the full sample.

Evaluations of Ambiguous Evidence

To examine whether police evaluations of ambiguous evidence aligned with the Bayesian Cognitive Model or confirmation bias, we analyzed a 2 (social norms: thorough vs. efficient) x 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) between-groups factorial ANOVA on evaluations of ambiguous evidence. Neither the norms manipulation, evidence order, nor their interaction was significant (ps > 0.60).

However, there was a significant interaction between the norms manipulation and evidence order once removing participants who reported they did not pay attention or who failed to correctly remember the manipulation, $F(1, 38) = 4.56$, $p = 0.04$, $\eta_p^2 = 0.11$. Surprisingly, planned contrasts revealed that evaluations of ambiguous evidence were significantly more incriminating when participants encountered ambiguous evidence before incriminating evidence ($M = 5.23$, $SD = 1.33$) than when participants encountered

incriminating evidence before ambiguous evidence (M = 4.83, SD = 0.97), but only in the efficient condition $t(38) = 2.65$, $p = .03$, $d = 1.07$, 95% CI [0.13, 1.98], and not in the thoroughness condition ($p = .54$). Thus, our findings did not support the Bayesian Cognitive Model or confirmation bias.

Study 1 Discussion

We would expect final guilt judgments to be the same across order conditions if police were updating their beliefs according to the Bayesian Cognitive Model, but instead we found police participants to be biased in their final guilt judgments. Police participants exhibited a recency bias, rather than a confirmation bias, such that they reported higher guilt judgments when they encountered ambiguous evidence first compared to when they encountered the incriminating evidence first. In other words, police participants' final guilt judgments were most aligned with the last piece of evidence police encountered, which is consistent with past research (e.g., Charman et al., 2016; Dahl et al., 2009)

However, we only found a recency bias effect in final guilt judgments when police participants were primed with social norms of efficiency. Social norms of thoroughness eliminated this recency bias. Thus, social norms of thoroughness led police to judge suspect guilt as more Bayesian, whereas social norms of efficiency led to a recency bias. Ask and colleagues (2011) suggested norms of thoroughness led police to process later evidence more when compared to efficiency norms, whereas efficiency norms led police to be more likely to discount evidence discovered later in an investigation. Our findings suggest norms of thoroughness might lead to greater processing of all evidence by minimizing over-weighing the last piece of evidence.

According to the Bayesian Cognitive Model, we would also expect police

participants in the incriminating first-ambiguous second condition to have similar initial and final guilt judgments because the ambiguous evidence should offer no new information. Our results suggest police participants updated their guilt judgments according to the Bayesian Cognitive Model, as there was no difference between initial and final guilt judgments in the incriminating evidence first condition. However, it is possible this null finding occurred due to our small police sample.

Police participants' initial guilt judgments were significantly lower than their final guilt judgments when they encountered ambiguous evidence first and incriminating evidence second, which could suggest police are updating their guilt beliefs according to the information provided by the incriminating evidence. However, our results suggest police participants were over-weighing the incriminating evidence because we found order affected final guilt judgments. We did find that social norms of thoroughness weakened the extent to which police participants weighed the incriminating evidence when updating their final guilt judgments, which could suggest thoroughness norms led to police being more conservative when updating their beliefs.

Our findings were the same in the full and reduced dataset when examining guilt judgments but not when examining evaluations of ambiguous evidence. There were no experimental effects of evidence order or social norms on evaluations of ambiguous evidence when analyzing the full dataset, which could suggest police participants' evaluations were aligned with the Bayesian Cognitive Model. However, we did find a significant interaction between the norms manipulation and evidence order after removing police participants who failed the attention check and manipulation checks. Police participants evaluated the ambiguous evidence as more incriminating in the

ambiguous first-incriminating second condition than in the incriminating first-ambiguous second, but only when we primed police with norms of efficiency. This finding suggests police participants in the thoroughness norms condition were Bayesian in their evaluations of ambiguous evidence.

However, our data suggest police in the efficiency norms condition were not evaluating evidence according to the Bayesian Cognitive Model or confirmation bias as we predicted. Rather, our data might suggest a contrast effect because police evaluated the ambiguous evidence as *less* incriminating after encountering incriminating evidence (vs. before encountering incriminating evidence). Perhaps police in the incriminating first-ambiguous second condition compared the ambiguous evidence to the incriminating evidence, thereby rating the evidence as less incriminating compared to police who had no other reference point.

Study 2

Study 1 was limited by its sample size and it could not provide information as to whether police judgments differ from the general population. Study 2 addresses these gaps. The design is identical to Study 1, such that Study 2 employed a 2 (order: incriminating evidence first vs. incriminating evidence last) x 2 (norms: efficiency vs. thoroughness) design and counterbalanced evidence type to account for the potential confound between evidence order and evidence type.

Study 2 Method

Participants

We solicited participants from Amazon Mechanical Turk (MTurk), which is an online labor pool, to participate in our study for a $1.00 reward. This solicitation was

only available to workers with IP addresses in the United States who had completed 1000 previous assignments and had at least a 98% approval rate. After recruiting 636 participants, we conducted the following validity checks. First, we asked participants to copy and paste a randomized code from the end of our survey into the MTurk website to verify they completed the entire study. Second, we employed a two-part validity check by asking participants to select their date of birth (day, month, year) from drop-down menus at the beginning of the study and to type their age at the end of the study (Cobanoglu et al., 2021). We removed participants whose self-reported age deviated by more than 1 year of their self-reported year of birth. Third, we embedded a hidden question into the survey to detect bots. Fourth, the first author read answers to the open-ended questions and removed participants who 1) copy and pasted text from the study, 2) wrote answers that did not answer the question or indicated poor comprehension (e.g., writing "good" or "nice"). Fifth, we removed participants who did not complete the suspicion check, manipulation check, and demographic questions. The remaining participants (N = 525) self-identified their gender as 57% men, 42% women, and 0.6% other. Participants' self-identified race and ethnicity were 76% Caucasian, 8% African American, 4% Latino/a, 9% Asian, 2% Native American, and 2% multi-ethnic. The mean age was 40.89 (SD = 12.33; range = 19-71). According to an α of .05, β of .80, and an assumed effect size of η_P^2= 0.06 based on the smallest effect size found in Study 1 (f = 0.253), the required sample size was 125. Thus, this study recruited a sufficient sample size to detect our hypothesized effects.

Design, Materials, and Procedure

The design, materials, and procedure were the same as in Study 1, with the

following exceptions. First, participants entered their MTurk Worker ID at the beginning of the study. Second, participants answered their date of birth (day, month, year) from drop-down menus at the beginning of the study. Third, a hidden question was embedded into the survey to detect bots. Fourth, the online instructions for the norms manipulations indicated a large majority (i.e., more than 80%) of previous survey responders (vs. police officers) agreed with the same list of six statements regarding the quality of a good police investigator (Ask et al., 2011). Last, we compensated participants $1.00 after completing the study.

Results

We conducted the same analyses as in Study 1 using R.

Attention, Suspicion, and Manipulation Checks

One participant reported they either did not read or pay attention to the study. No participants correctly inferred the purpose of the norms manipulation or that the order of evidence varied. The manipulation checks revealed 41 participants incorrectly remembered the norms manipulation, 95 participants incorrectly remembered the order they encountered the evidence, and 48 participants incorrectly remembered which type of evidence was incriminating. We performed the analyses with (N = 525) and without the data (N = 392) from the participants who reported they did not pay attention or who failed to correctly remember the manipulation. The pattern of results was the same between the full and reduced sample for updating guilt judgments, but not for final and ambiguous guilt judgments.

Preliminary Analyses

First, participants' agreement with the norms statements again differed between

the thoroughness and efficiency conditions $t(523) = 12.24$, $p < .001$. Participants were more likely to agree with the thoroughness statements ($M = 7.98$, $SD = 1.17$) than the efficiency statements ($M = 6.61$, $SD = 1.38$). Unlike Study 1, the ratings did significantly correlate with several of the main dependent variables and they qualified the analyses on how participants updated their guilt judgments, and therefore their agreement with these norms statements were included in the analyses as a covariate.

Second, we examined whether the effects of the norms manipulation or evidence order on the main dependent variables depended upon evidence type to determine whether these effects are confounded by evidence type. Evidence type significantly moderated our experimental conditions on how participants updated their guilt judgments (ps < .001), so evidence type was included as a control for this analysis. Evidence type did not interact with evidence order or the norms manipulation on final guilt judgments or evaluations of ambiguous evidence (ps > 0.06) thus was not included in these analyses.

Final Guilt Judgments

As in Study 1, we examined the effects of our manipulations on final guilt judgments to identify whether evidence order affected guilt judgments according to the Bayesian Cognitive Model, confirmation bias, or recency bias, and whether social norms moderated the effects of bias. Specifically, we analyzed a 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (social norms: thorough vs. efficient) between-groups factorial ANCOVA on final guilt judgments with the agreement to the norms statements entered as a covariate. There were no significant main effects of evidence order or the norms manipulation on final guilt judgments (ps > .08). The interaction between our manipulations was not significant and

planned univariate analyses did not reveal any significant patterns (*p*s > .19).

After we removed participants who reported they did not pay attention or who failed to correctly remember the manipulation, the norms manipulation was significant, $F(1, 387) = 6.39$, $p = .01$, $\eta_p^2 = 0.02$. Participants' final guilt judgments were higher in the efficient norms condition ($M = 6.23$, $SD = 1.59$) than the thoroughness norms condition ($M = 6.06$, $SD = 1.55$). This effect suggests MTurk participants were more conservative in their guilt judgments in the thoroughness condition than in the efficiency condition. The main effect of evidence order was still not significant ($p = .07$).

Although the two-way interaction between evidence order and social norms did not reach conventional levels of significance, $F(1, 387) = 3.05$, $p = 0.36$, $\eta_p^2 = 0.002$, we conducted the same planned analyses as in Study 1. Specifically, participants' final guilt judgments approached being significantly higher when they encountered ambiguous evidence before incriminating evidence ($M = 6.60$, $SD = 1.58$) than when participants encountered incriminating evidence before ambiguous evidence ($M = 6.17$, $SD = 1.49$) in the efficient norms condition, $t(387) = 1.96$, $p = .05$, $d = 0.28$, 95% CI [0.00, 0.87]. There was no difference between evidence order in the thorough norms condition ($p = .51$). Thus, MTurk participants trended toward exhibiting a recency bias in their judgments of suspect guilt, but only when primed with norms of investigative efficiency.

Updating Guilt Judgments

As in Study 1, we used a linear mixed-effects model to examine whether initial and final guilt ratings differed by evidence order to determine whether police were Bayesian when updating their beliefs or if they displayed a confirmation or recency bias. We also used this model to examine whether social norms moderated biased effects,

should any occur. Fixed effects included the norms manipulation, evidence order manipulation, guilt judgment time (i.e., initial vs. final), their interactions, evidence type, and agreement to the norms statements. Random effects included the participants' ID.

There was a significant effect of norms statements, $F(1, 519) = 5.20$, $p = .02$, guilt judgment time, $F(1, 521) = 8.51$, $p = .004$, and evidence order, $F(1, 519) = 31.99$, $p < .001$, which were qualified by a significant interaction between guilt judgment time and evidence order $F(1, 521) = 179.88$, $p < .001$. Planned contrasts revealed that initial guilt judgments were significantly lower ($M = 5.23$, $SD = 1.54$) than final guilt judgments ($M = 6.31$, $SD = 1.64$) when participants encountered the ambiguous evidence before incriminating evidence $t(521) = 11.55$, $p < .0001$, $d = 1.01$, 95% CI [0.90, 1.26], suggesting participants were updating their beliefs in response to the incriminating evidence. Conversely, initial guilt judgments were significantly higher ($M = 6.78$, $SD = 1.48$) than final guilt judgments ($M = 6.08$, $SD = 1.67$) when participants encountered incriminating evidence first $t(521) = 7.42$, $p < .0001$, $d = 0.65$, 95% CI [0.51, 0.88], suggesting participants were decreasing their guilt beliefs in response to the ambiguous evidence.

Although social norms only marginally moderated the interaction between guilt judgment type and evidence order $F(1, 521) = 3.15$, $p = .08$, planned univariate analyses revealed a similar pattern to Study 1. Participants who encountered ambiguous evidence first and incriminating evidence second had lower initial guilt judgments ($M = 5.27$, $SD = 1.50$) than final guilt judgments ($M = 6.47$, $SD = 1.68$) when primed with efficient norms, $t(521) = 9.37$, $p < .0001$, $d = 1.12$, 95% CI [0.95, 1.45]. This effect was *weaker* when participants were primed with thorough norms, such that their initial guilt judgments

were (M = 5.19, SD = 1.57) significantly lower than their final guilt judgments (M = 6.15, SD = 1.60) when they encountered ambiguous evidence first and incriminating evidence second, $t(521) = 7.04$, $p < .001$, $d = 0.90$, 95% CI [0.69, 1.23].

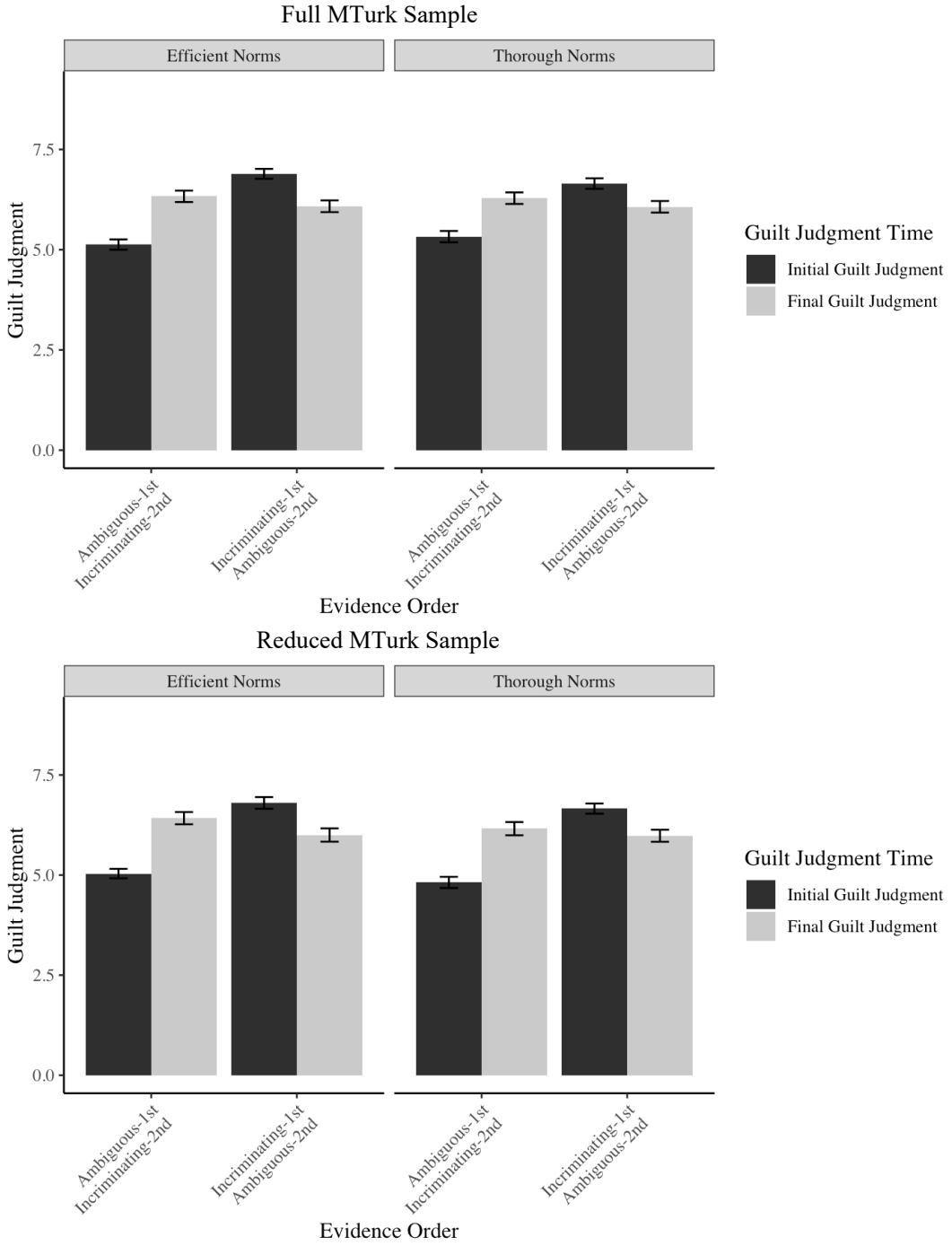
Unlike Study 1, there was also an interaction between the norms statements and guilt judgment time among participants who encountered incriminating evidence first and ambiguous evidence second, such that participants' initial guilt judgments (M = 7.03, SD = 1.36) were significantly higher than their final guilt judgments (M = 6.22, SD = 1.60) when primed with efficient norms, $t(521) = 5.86$, $p < .0001$, $d = 0.76$, 95% CI [0.54, 1.08]. This effect suggests a recency bias. This effect was *weaker* when participants were primed with thorough norms such that their initial guilt judgments were (M = 6.52, SD = 1.57) significantly higher than their final guilt judgments (M = 5.94, SD = 1.73) when they encountered incriminating evidence first, $t(521) = 4.59$, $p < .001$, $d = 0.54$, 95% CI [0.33, 0.83]. Together, these findings suggest that social norms of thoroughness led MTurk participants to be more conservative when updating their guilt judgments thereby decreasing their recency bias (see Figure 2).

After removing participants who reported they did not pay attention or who failed to correctly remember the manipulation, the pattern of effects remained the same. There was still a significant main effect of norms statements on total guilt judgments $F(1, 388) = 9.95$, $p = .001$ in the same direction as the previous analyses. There were also still significant main effects of guilt judgment time $F(1, 388) = 18.46$, $p < .001$ and evidence order $F(1, 386) = 44.92$, $p < .001$ that was qualified by a significant interaction between guilt judgment time and evidence order $F(1, 388) = 210.95$, $p < .001$. This interaction effect also followed the same pattern as the full sample. Finally, although the three-way

interaction was still not significant ($p = 0.57$), the univariate analyses revealed the same pattern we reported for the full sample (see Figure 2).

Figure 2

Effects of Evidence Order and Social Norms on MTurk Worker's Guilt Judgments



Note. The above figure depicts how the order MTurk participants encountered evidence and social norms influenced how they updated their guilt judgments in the full and reduced samples. As in Study 1, All initial guilt judgments were in response to the first

piece of evidence, which differed in strength depending upon order condition. All final guilt judgments were after receiving both pieces of evidence. Participants reported their guilt judgments on a 1-9 scale, such that answers above 5 suggest guilt, below 5 suggest innocence, and 5 suggest "*neither guilty nor innocent*."

Evaluations of Ambiguous Evidence

To examine whether participants' evaluations of ambiguous evidence aligned with the Bayesian Cognitive Model or confirmation bias, we analyzed a 2 (social norms: thorough vs. efficient) x 2 (evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) between-groups factorial ANCOVA on evaluations of ambiguous evidence with the agreement with the norms statements entered as a covariate. There was a significant main effect of the norms manipulation on evaluations of ambiguous evidence $F(1, 520) = 4.05$, $p = 0.04$, $\eta_p^2 = 0.008$. Participants evaluated the ambiguous evidence as significantly less incriminating in the efficient norms condition ($M = 4.81$, $SD = 1.47$) than in the thoroughness norms condition ($M = 5.07$, $SD = 1.55$). However, this effect did not appear when analyzing the data without participants who reported they did not pay attention or who failed to correctly remember the manipulation ($p = .25$). There were no other significant effects after removing participants who did not pay attention or failed to correctly remember the manipulation.

Study 2 Discussion

We found MTurk participants in the reduced sample trended toward a recency bias in their final guilt judgments, such that their final guilt judgments were marginally higher when they encountered the ambiguous evidence first than when they encountered the incriminating evidence first. This trend is similar to Study 1, suggesting a recency bias. However, we only found this trend among participants primed with norms of

efficiency, whereas norms of thoroughness minimized this effect.

Similar to Study 1, we also found MTurk participants' initial guilt judgments were significantly lower than their final guilt judgments when they encountered ambiguous evidence before incriminating evidence, but this effect was weaker when primed with norms of investigative thoroughness (vs. efficiency). This finding again suggests norms of thoroughness led participants to update their guilt judgments more conservatively than norms of efficiency. This finding also suggests participants in the ambiguous first-incriminating second condition who were primed with efficiency norms were over-weighing the incriminating evidence when considering that norms of efficiency led participants to display a recency bias in their final guilt judgments.

Unlike Study 1, MTurk participants in the incriminating first-ambiguous second condition did not update their guilt judgments according to the Bayesian Cognitive Model because their initial guilt judgments were higher than their final guilt. This finding also suggests a recency bias because participants were relying upon the ambiguous evidence when reporting their final guilt judgments. Social norms of thoroughness did weaken this effect when compared to social norms of efficiency, thus social norms of thoroughness did lead participants in the incriminating first-ambiguous second to be more Bayesian when updating their guilt judgments.

Unlike Study 1, we only found a main effect of social norms on evaluations of ambiguous evidence. MTurk participants evaluated the ambiguous evidence as less incriminating when we primed them with efficient norms than when we primed them with thorough norms. However, this effect disappeared after removing participants who did not pay attention or failed to correctly remember the manipulation.

General Discussion

This research had the goals to test competing hypotheses regarding the effects of evidence order on guilt judgments and evaluations of ambiguous evidence, as well as to examine a potential method to minimize bias through social norms of thoroughness. Consistent with past research, participants exhibited a recency bias in their guilt judgments (Charman et al., 2016; Dahl et al., 2009; Price & Dahl, 2013). Norms of thoroughness did minimize this bias and tended to make police and laypeople more conservative and Bayesian in their guilt judgments. Overall, the trend in findings was the same between samples for final guilt judgments, but there were differences between samples regarding how participants updated their guilt judgments and evaluated ambiguous evidence.

We originally predicted three competing hypotheses regarding participants' final guilt judgments. According to the Bayesian Cognitive Model, final guilt judgments should be the same between order conditions because people should evaluate each piece of information independently (Druckman & McGrath, 2019; Edwards, 1962), regardless of social norms. However, much of previous research found a recency effect that the last piece of information has the greatest effect on final guilt judgments compared to other information (e.g., Charman et al., 2016; Dahl et al., 2009), suggesting people are not Bayesian in their guilt judgments. Other research found the first piece of evidence predicted final guilt judgments (Charman et al., 2017), thus there is mixed research as to whether people over-weigh the first or last piece of information in final guilt judgments. If people are not Bayesian when forming their final guilt judgments, social norms of thoroughness (vs. efficiency) could potentially minimize this bias (Ask et al., 2011).

Participants in both samples who encountered ambiguous evidence before incriminating evidence rendered higher final guilt judgments than participants who encountered incriminating evidence before ambiguous evidence, suggesting police's final guilt judgments are subject to a recency bias. However, this effect only emerged when participants were in the efficient social norms condition and not in the thoroughness norms condition, which implies social norms of thoroughness could potentially eliminate recency bias.

We also presented three competing hypotheses regarding how participants could update their guilt beliefs after considering the second piece of evidence. According to the Bayesian Cognitive Model, participants in the incriminating evidence first condition should have the same initial and final guilt judgments because the ambiguous evidence does not offer any new information (Druckman & McGrath, 2019). Previous research found police displayed confirmation bias when their final guilt judgments were stronger than their initial guilt judgments after encountering additional ambiguous evidence (Charman et al., 2017), but a recency bias would predict the opposite in that the last piece of evidence would have the most weight on final guilt judgments (Charman et al., 2016). However, research examining the effects of evidence order on guilt judgments did not measure initial guilt beliefs (Charman et al., 2016; Dahl et al., 2009) or did not vary evidence order and compare initial guilt judgments to final guilt judgments (Charman et al., 2017).

The current study expanded upon these previous gaps. We found police participants who encountered incriminating evidence before ambiguous evidence rendered guilt judgments that aligned with the Bayesian Cognitive Model because their

initial and final guilt judgments were not statistically different. However, MTurk participants who encountered incriminating evidence before ambiguous evidence displayed a recency bias when updating their guilt judgments, although this effect was weaker in the thoroughness norms condition compared to the efficiency norms condition. Thus, police participants were more Bayesian than MTurk participants when updating their guilt judgments, whereas MTurk participants displayed a recency bias. Perhaps police's experience with investigations led them to incorporate ambiguous evidence more accurately into their guilt judgments.

We also found both police and laypeople who encountered ambiguous evidence before incriminating evidence increased their guilt judgments after encountering the incriminating evidence. Initially, this belief updating appears rational that participants updated their beliefs after considering the incriminating evidence. However, our data suggests participants over-weighed the incriminating evidence when considering our finding that evidence order affected final guilt judgments. Norms of thoroughness weakened the extent to which participants increased their guilt judgments after encountering the incriminating evidence, which suggests social norms of thoroughness led police and laypeople to be Bayesian in their guilt judgments by not over-weighing the incriminating evidence.

Finally, we offered two competing predictions regarding evaluations of ambiguous evidence. The Bayesian Cognitive Model would suggest evidence order would not affect evaluations of ambiguous evidence (Druckman & McGrath, 2019), whereas confirmation bias would suggest evaluations of ambiguous evidence would be more incriminating after police encounter incriminating evidence (Charman et al., 2017).

However, police participants evaluated ambiguous evidence as more incriminating when they encountered the evidence first (vs. second), but only in the efficiency condition. Police might have been comparing the ambiguous evidence to the incriminating evidence to produce somewhat of a contrast effect (see Price & Dahl, 2013 for an explanation of how evidence order could produce a contrast effect). Perhaps the ambiguous evidence appeared more strongly exculpatory compared to the incriminating evidence. Still, norms of thoroughness minimized this contrast effect. Surprisingly, MTurk participants evaluated ambiguous evidence as less incriminating in the efficient norms condition than in the thoroughness norms condition, though this effect disappeared after removing participants who failed the manipulation and suspicion checks. Therefore, we did not find support for either the Bayesian Cognitive Model or confirmation bias on evaluations of ambiguous evidence in either sample.

Despite different evaluations of ambiguous evidence between samples, the overall trend remained the same between samples: Norms of thoroughness minimized bias in all guilt judgments and evidence evaluations when compared to norms of efficiency, regardless of the sample. Thus, despite the inconsistent findings on how participants update their guilt judgments when they encounter incriminating evidence before ambiguous evidence and on their evaluations of ambiguous evidence, we argue our results still have direct implications for the criminal justice system.

Based on our findings, we echo the conclusions and recommendations from Ask and colleagues (2011). Police guilt judgments were more likely to be biased when efficiency norms were salient, which means innocent people who become suspects during a criminal investigation could be at risk to be falsely incriminated if police are incorrectly

weighing evidence in their final guilt judgments. For example, police who first encounter ambiguous or exculpatory evidence and later encounter incriminating evidence might over-weigh the last piece of incriminating evidence, thereby falsely incriminating an innocent suspect. Police departments should work on promoting norms of thoroughness, rather than efficiency, during investigations to increase the likelihood of police properly judging suspect guilt.

As with any study, there are limitations. Notably, Study 1 recruited a small sample of police. Relatedly, there was also a low response rate from the police departments, as only 4 departments out of 160 contacted agreed to participate. Because police are a hard-to-reach population, future research should employ a longer recruitment period, contact more departments, and build rapport with more departments as potential methods to increase sample sizes.

To account for police being a hard-to-reach sample, future research can also recruit from easier-to-reach populations, such as those from MTurk or college students to act as a comparison group to police officers. Perhaps it is possible to generalize findings from other populations (e.g., MTurk workers, college students) if patterns of results are similar between laypeople and officers. For example, similar trends were found previously among police and student samples (Charman et al., 2017; Yang et al., 2022). Should research continue to produce similar trends between police and other convenience samples, then we can assume findings from samples other than police are generalizable which can further research on police decision-making. This approach is not novel, as jury researchers have used convenience samples to infer how actual jurors make decisions. A meta-analysis has even suggested mock jurors from convenience samples made similar

judgments to real jurors (Bornstein et al., 2017). Thus, future research should recruit from both laypeople samples and police to determine the extent to which findings from convenience samples are generalizable to police judgments.

Another limitation is that many participants in the police sample (32%) and the laypeople sample (25%) failed at least one manipulation memory check question or attention check question. Surprisingly, 13% of the police sample reported they did not pay attention or read the study materials, although only one participant on MTurk reported not paying attention or reading the study materials. Perhaps MTurk workers were less likely to explicitly report their attention because their compensation relies upon passing attention checks.

There could be a reason to believe our manipulation memory check questions were confusing because 25% of participants in both samples failed at least one manipulation memory check. Given the data collected, it is unclear whether participants could not recall the manipulations or did not comprehend the manipulation check questions. Most participants who failed the manipulation memory check in the police sample ($N = 11$, 50%) and the MTurk sample ($N = 95$, 71%) did not correctly recall the order, suggesting participants either did not consciously encode the order they encountered the evidence, or they did not understand the manipulation check question. People tend to remember the order of information presented in an auditory format better than in a visual format (Unnava et al., 1994), so the visual nature of our materials might have impeded conscious recollection of the evidence order. Another possibility is that perhaps the differences in results between full and reduced samples could be attributed to a difference in people who were consciously aware of the evidence order and those who

were unconsciously aware, but still affected, by evidence order.

Of course, laboratory experiments are inherently limited in their scope of ecological validity, which applied to our study in three major ways. First, we only included two pieces of evidence to examine order effects. Police could realistically encounter many more pieces of evidence, and thus our findings might not be generalizable to investigations with more than two pieces of evidence. Second, our study used written descriptions of case scenarios and evidence, which does not reflect an actual police investigation that might include physical evidence or working with other officers. Third, there was a lack of consequentiality in our experiment. Police involved in real investigations must use evidence and judge a suspect's guilt to make subsequent decisions, whereas in our study there were no consequences to evaluating the evidence or judging the suspect's guilt.

A final limitation is that this study was not a true test of the Bayesian Cognitive Model because we used classic hypothesis tests. It is possible our non-significant findings in the thoroughness norms condition were due to a lack of power, rather than participants judging suspect guilt in a Bayesian manner. Although a lack of power could be a reasonable explanation regarding the police sample findings, the laypeople sample was adequately powered and still showed that thoroughness norms eliminated bias. Still, future research could use equivalence testing as an alternative analysis plan to examine whether police are Bayesian in their judgments, as equivalence testing can examine whether two means are equivalent (versus hypothesis testing that calculates whether two means are different). Future research could also use Bayesian statistical analyses to examine such effects.

Conclusion

Minimizing biases in police investigations can ensure the correct perpetrator is incriminated and avoid wrongfully convicting innocent people such as Christopher Tapp. Because police investigators evaluate multiple pieces of evidence during an investigation, it is critical to not only understand factors that bias evaluations of evidence and guilt judgments, but also factors that can minimize biases. We found evidence order can result in a recency bias such that police over-weigh the last piece of evidence they encounter, but promoting norms of investigative thoroughness (vs. efficiency) during an investigation can minimize this bias. These findings provide an evidence-based method to promote effective investigations to ensure correct suspects are incriminated and innocent suspects are not wrongfully incriminated.

References

Ask, K., & Alison, L. (2010). Investigators' decision-making. In P. A. Granhag (Ed.), *Forensic Psychology in Context* (1st ed., pp. 35–55). Willan. https://doi.org/10.4324/9781315094038-3

Ask, K., Granhag, P. A., & Rebelius, A. (2011). Investigators under influence: How social norms activate goal-directed processing of criminal evidence. *Applied Cognitive Psychology*, *25*(4), 548–553. https://doi.org/10.1002/acp.1724

Blair, J. P., & Rossmo, D. K. (2010). Evidence in context: Bayes' theorem and investigations. *Police Quarterly, 13*, 123–135.10.1177/1098611110365686

Bornstein, B. H., Golding, J. M., Neuschatz, J., Kimbrough, C., Reed, K., Magyarics, C., & Luecht, K. (2017). Mock juror sampling issues in jury simulation research: A meta-analysis. *Law and Human Behavior*, *41*(1), 13–28. https://doi.org/10.1037/lhb0000223

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. https://doi.org/10.1037/a0026450

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, *7*(2), 91–103. https://doi.org/10.1037/1076-898X.7.2.91

Charman, S. D., Carbone, J., Kekessie, S., & Villalba, D. K. (2016). Evidence evaluation and evidence integration in legal decision-making: Order of evidence presentation as a moderator of context effects. *Applied Cognitive Psychology*, *30*(2), 214–225. https://doi.org/10.1002/acp.3181

Charman, S. D., Gregory, A. H., & Carlucci, M. (2009). Exploring the diagnostic utility of facial composites: Beliefs of guilt can bias perceived similarity between composite and suspect. *Journal of Experimental Psychology: Applied*, *15*(1), 76–90. https://doi.org/10.1037/a0014682

Charman, S. D., Kavetski, M., & Mueller, D. H. (2017). Cognitive bias in the legal system: Police officers evaluate ambiguous evidence in a belief-consistent manner. *Journal of Applied Research in Memory and Cognition*, *6*(2), 193–202. https://doi.org/10.1016/j.jarmac.2017.02.001

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cobanoglu, C., Cavusoglu, M., & Turktarhan, G. (2021). A beginner's guide and best practices for using crowdsourcing platforms for survey research: The Case of Amazon Mechanical Turk (MTurk). *Journal of Global Business Insights*, *6*(1), 92–97. https://doi.org/10.5038/2640-6489.6.1.1177

Dahl, L. C., Brimacombe, C. A. E., & Lindsay, D. S. (2009). Investigating investigators: How presentation order influences participant–investigators' interpretations of eyewitness identification and alibi evidence. *Law and Human Behavior*, *33*(5), 368–380. https://doi.org/10.1007/s10979-008-9151-y

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in

climate change preference formation. *Nature Climate Change*, *9*(2), 111–119. https://doi.org/10.1038/s41558-018-0360-1

Edwards, W. (1962). Dynamic decision theory and probabilistic information processings. *Human Factors*, *4*(2), 59–74. https://doi.org/10.1177/001872086200400201

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In Ron Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge University Press.

Hogg, M. A. (2010). Influence and leadership. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470561119.socpsy002031

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *WIREs Cognitive Science*, *2*(1), 8–21. https://doi.org/10.1002/wcs.80

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, *27*(2), 187–203. https://doi.org/10.1023/A:1022599230598

Murphy, H. (2019, July 18). The jury said he killed her daughter. She helped clear his name. *The New York Times.* Retrieved, https://www.nytimes.com/2019/07/18/us/angie-dodge-christopher-tapp.html

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

Otterbourg, K. (2021, February 10). Christopher Tapp. *The National Registry of Exonerees.* Retrieved,

https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=5592

Price, H. L., & Dahl, L. C. (2014). Order and strength matter for evaluation of alibi and eyewitness evidence. *Applied Cognitive Psychology*, *28*(2), 143–150. https://doi.org/10.1002/acp.2983

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

Scherr, K. C., Redlich, A. D., & Kassin, S. M. (2020). Cumulative disadvantage: A psychological framework for understanding how innocence can lead to confession, wrongful conviction, and beyond. *Perspectives on Psychological Science*, *15*(2), 353–383. https://doi.org/10.1177/1745691619896608

Shah, J. (2003). Automatic for the people: How representations of significant others implicitly affect goal pursuit. *Journal of Personality and Social Psychology*, *84*(4), 661–681. https://doi.org/10.1037/0022-3514.84.4.661

Snook, S. A. (2000). *Friendly fire: The accidental shootdown of U.S. Black Hawks over Northern Iraq*. Princeton University Press.

The Professional Research Pool for Criminal Justice Science. (n.d.). https://www.prpforcjscience.org/

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. (2017). *Law enforcement agency roster (LEAR), 2016: Version 1* [Data set]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR36697.V1

Unnava, H. R., Burnkrant, R. E., & Erevelles, S. (1994). Effects of presentation order and communication modality on recall and attitude. *Journal of Consumer Research*, *21*(3), 481–490. https://doi.org/10.1086/209412

Yang, Y., Madon, S., Cabell, J. J., Moody, S. A., Guyll, M. (2022). *The effect of presumption of guilt on police guilt judgments*. Unpublished Manuscript.

Appendix

Norms Manipulations (adapted from Ask et al., 2011)
This survey contains questions that have been used in surveys [of police officers (Study 1)] in the past. Previously, a large majority (more than 80%) [of officers (Study 1)] agreed with these statements. We are interested in how many [officers (Study 1)/ people (Study 2)] agree with these statements today.
Please rate the extent to which you agree with each statement from 1 (*strongly disagree*) to 9 (*strongly agree*).
All Conditions
1. A good investigator knows how to make good use of his/her prior practical experience.
2. A good investigator has good communication skills.
Efficiency Norm Condition
3. A good investigator is decisive.
4. A good investigator should make quick inferences from complex material.
5. A good investigator often sees a solution to a crime early in the investigation.
6. A good investigator is focused on minimizing the amount of time spent on a case unnecessarily.
Thoroughness Norm Condition
3. A good investigator is patient and systematic.
4. A good investigator should avoid premature conclusions about a crime.
5. A good investigator does not let his/her first impression of a case alter his/her view.
6. A good investigator would rather spend extra time working on a case than fail to investigate an important detail.

Case Scenario (adapted from Charman et al., 2016)
On June 19th, a teenaged girl named Alexis Schafman was found dead in a bush. Alexis was attending her brother's soccer game at a park in Colorado in the early evening. During a break in the game, she told her family that she was going to "walk around for a bit." When she did not return, her family grew concerned and her father began to search for her. After the game was over and she had not returned or been found, the Schafmans called the police. The police performed an initial search of the nearby woods and were unsuccessful at finding her. However, the police returned to search a second time later in the evening and they performed a canine-assisted search of the woods. The dogs led them to a bush in which Alexis's body had been thrown. She was declared dead at the scene. It was clear that the cause of death was strangulation and there appeared to have been an intense struggle. The investigation continued to narrow for about 2 weeks until police settled on a suspect, Samuel Scott, who matched a general description given by a witness.

Evidence Manipulations
Incriminating DNA Evidence
The forensic expert determined the characteristics of the DNA sample from the crime scene are 100,000 times more likely if the DNA sample came from the suspect, Samuel Scott, than if the sample came from a randomly chosen person.

Incriminating Eyewitness Evidence
The investigative expert determined that the characteristics of the eyewitness ID evidence are 100,000 times more likely if the suspect, Samuel Scott, was at the crime scene than if the suspect was a randomly chosen person.

Ambiguous DNA Evidence
The forensic expert determined it is inconclusive as to whether the DNA sample from the crime scene matched the suspect, Samuel Scott, so the DNA evidence provides no support for whether Samuel Scott was at the crime scene or not at the crime scene.

Ambiguous Eyewitness Evidence
The investigative expert determined that the eyewitness was not sure whether the suspect, Samuel Scott, was in the lineup, so the eyewitness identification evidence provides no support for whether Samuel Scott was at the crime scene or not at the crime scene.

Suspicion, Attention, and Manipulation Check

4. Please indicate what you knew about this study before participating. [open response]
5. Do you think the purpose of this study is obvious?
    a. Yes
    b. No
6. Please indicate what research questions you believe might be under investigation in this study.

--Page Break--

3. Did you pay attention to this study?
4. Did you read the full case scenario?

--Page Break--

2. Which series of statements do you remember reading from study 1?
    a. Option A:
        1. A good investigator is decisive.
        2. A good investigator has the ability to make quick inferences from a complex material.
        3. A good investigator often solves a crime early in the investigation.
        4. A good investigator is focused on minimizing the amount of time spent on a case unnecessarily.
    b. Option B
        1. A good investigator is patient and systematic.
        2. A good investigator has the ability to avoid premature conclusions about a crime.
        3. A good investigator does not let his/her first impression of a case color his/her view.
        4. A good investigator would rather spend extra time working on a case than fail to investigate an important detail.
    c. I do not remember.
3. In the second study, which set of statements best describes the types of evidence

you read about?

a. 1."The forensic expert concluded the characteristics of the DNA sample from the crime scene were 100,000 times more likely if the DNA sample came from the suspect than if the sample came from a randomly chosen person."
   2."The investigative expert concluded the eyewitness was not sure whether the suspect was in the lineup, so the eyewitness identification evidence provides no support for whether the suspect was at the crime scene or not at the crime scene."
b. 1."The forensic expert concluded it is inconclusive as to whether the DNA sample from the crime scene matched the suspect, so the DNA evidence provides no support for whether the suspect was at the crime scene or not at the crime scene."
   2."The investigative expert concluded the characteristics of the eyewitness ID evidence were 100,000 times more likely if the suspect was at the crime scene than if the suspect was a randomly chosen person."
c. I do not remember.

4. In the second study, what order did you encounter the two pieces of evidence?
   a. DNA evidence first, eyewitness identification evidence second
   b. Eyewitness identification evidence first, DNA evidence second
   c. I do not remember

Demographic Questionnaire

*The first three questions were only asked in the police sample (Study 1)

11. How many years of experience as a police officer do you have? [open response]
12. What type of law enforcement agency do you currently work for?
    a. Local
    b. State
    c. Federal
    d. Other (please describe)
13. What is your current ranking in your department?
    a. Chief of Police
    b. Deputy Chief
    c. Detective/investigator
    d. Patrol Officer
    e. Lieutenant
    f. Police Officer
    g. Sheriff
    h. Sergeant
    i. Captain
    j. Other (please describe)
14. What is your gender?
    a. Female
    b. Male
    c. Other

15. What is your age? ___
16. Please indicate your ethnicity/race:
    a. Black or African American
    b. Asian
    c. White or Caucasian
    d. Latina/o
    e. Native American
    f. Indian
    g. Multi-ethnic (Please indicate your ethnicity/race.) ____
    h. Other (Please indicate your ethnicity/race.) ____
17. What is your highest level of education?
    a. Some high school, no diploma
    b. High school graduate, diploma or the equivalent (for example: GED)
    c. Some college credit, no degree
    d. Trade/technical/vocational training
    e. Associate degree
    f. Bachelor's degree
    g. Master's degree
    h. Professional degree
    i. Doctorate degree
18. Generally speaking, would you describe your political views as…
    a. Very Liberal
    b. Somewhat Liberal
    c. Moderate
    d. Somewhat Conservative
    e. Very Conservative
19. Generally speaking, would you describe your political party as…
    a. Republican
    b. Democrat
    c. Independent
    d. Other ____
20. Do you have any questions or comments for the research team?

Chapter 5: Discussion and Conclusion

The purpose of this dissertation was to present a body of research regarding how police investigators evaluate evidence to judge suspect guilt. Paper #1 identified a gap in the literature regarding operationalizing evidence strength, proposed a definition based on competing hypotheses, provided nuance between subjective and objective evidence strength, and presented likelihood ratios (LRs) and logarithm likelihood ratios (Log LRs) as a method to define objective evidence strength. Paper #2 found police more accurately interpreted evidence strength presented in an LR format than a random match probability (RMP) format when judging suspect guilt, but only for DNA evidence. The presentation format of evidence strength did not affect judgments of suspect guilt for fingerprint evidence or eyewitness ID evidence. Paper #3 found participants were subject to a recency effect in their guilt judgments when evaluating multiple pieces of evidence, but social norms that promoted thorough investigations minimized this bias when compared to social norms that promoted efficient investigations.

The results from these three papers can help researchers and the legal system identify how accurate police investigators are in using evidence to form their judgments of a suspect's guilt, which ultimately has implications for preventing wrongful convictions. Because a wrongful conviction begins with police misidentifying the culprit of a crime, it is of critical importance to identify where these errors occur and methods to minimize these errors. These papers uniquely, yet collectively, added to the existing psychological literature on how police investigators evaluate evidence and how evidence affects police judgments of suspect guilt, while also providing empirically supported recommendations for police. This discussion section first provides a summary of each

paper, then integrates findings from each paper into primary conclusions, provides strengths and limitations of the three papers holistically, and concludes with avenues for future research.

Summary of Each Paper

Paper #1

Paper #1 identified that some research defined evidence strength as *the extent to which the evidence is incriminating* but this definition lacks an explicit alternative hypothesis. To explicate an alternative positively stated hypothesis, evidence strength should be defined as *the extent to which the evidence supports one hypothesis versus another hypothesis*. Such a definition can aid police for several types of evidence, regardless of whether the evidence is used to support hypotheses of guilt versus innocence or that a suspect left a fingerprint versus a randomly chosen person.

Separating subjective evidence strength from objective evidence strength into two different constructs can aid in determining how accurate police are in evaluating evidence strength. I defined subjective evidence strength as the evidence strength based on individual perceptions, whereas I defined objective evidence strength as the evidence strength based on a verifiably observed truth. I also proposed that LR and Log LRs can be used to precisely measure and communicate evidence strength. I concluded by identifying future directions, including that research should establish objective evidence strength for more types of evidence, identifying instruments to measure subjective evidence strength, and comparing subjective evidence strength to objective evidence strength to determine the extent to which police are accurate in evaluating evidence to judge suspect guilt.

Paper #2

Paper #2 examined whether the format to convey evidence strength and evidence type affected police accuracy in judging suspect guilt. I recruited police participants using CloudResearch and randomly assigned them to a 3 (evidence strength format: LR vs. RMP vs. neutral) x 3 (evidence type: DNA vs. fingerprint vs. eyewitness identification). I predicted police would be most accurate at judging suspect guilt when the evidence strength was presented in an RMP format (vs. LR), DNA (vs. fingerprint and eyewitness ID), and fingerprint (vs. eyewitness ID).

Surprisingly, I found police were most accurate in judging suspect guilt in the LR condition and that police in the RMP condition under-weighed the evidence by evaluating it the same as neutral evidence. However, this effect of evidence strength format was only found among police in the DNA evidence condition.

Although there were no main effects of evidence type on accuracy as predicted, there were differences between evidence types that were contingent upon evidence strength format. I did find that police significantly under-weighed fingerprint evidence compared to DNA evidence in the LR format condition. However, police were similarly accurate when looking at their mean scores on the accuracy measure; they were simply accurate in opposite directions.

Finally, I found that police rendered different guilt judgments between the log-scale guilt measure and the percent guilt measure. For example, police judged the suspect as more guilty when encountering DNA evidence than fingerprint evidence on the log-scale, whereas they judged the suspect as less guilty when encountering DNA evidence than fingerprint evidence on the percent guilt scale. Police also did not translate their

numeric responses on the log-scale to the percent guilt measure. That is, police's log-scale answers suggested a higher likelihood of guilt than their percent guilt measures. Thus, police did not render similar guilt judgments between the two guilt measures.

Paper #3

Paper #3 examined whether social norms and evidence order affected judgments of suspect guilt and evaluations of evidence. I recruited police participants using a hybrid snowball method and I recruited laypeople via Amazon Mechanical Turk (MTurk). I randomly assigned participants to a 2 (evidence order: evidence order: incriminating first-ambiguous second vs. ambiguous first-incriminating second) x 2 (norms: efficiency vs. thoroughness) between-subjects factorial design and counterbalanced evidence type within the evidence order condition. I predicted two competing hypotheses based on the Bayesian Cognitive Framework and confirmation bias. According to the Bayesian Cognitive Framework, I predicted that evidence order should not affect guilt judgments or evaluations of ambiguous evidence. According to confirmation bias, I predicted the initial piece of evidence should bias final guilt judgments in the direction of the first piece of evidence, so encountering incriminating evidence before ambiguous evidence would increase how incriminating police evaluated ambiguous evidence. I predicted biased effects would decrease when priming police with social norms of investigative thoroughness (vs. efficiency).

My findings supported a recency bias for police and laypeople's final guilt judgments, rather than the Bayesian Cognitive Framework or confirmation bias, but only when primed with efficient social norms and not when primed with thorough social norms. Police updated their guilt judgments according to the Bayesian Cognitive

Framework when they encountered incriminating evidence before ambiguous evidence, although the small sample size could have contributed to this non-significant effect. Police over-weighed incriminating when they encountered ambiguous evidence before incriminating evidence, which suggests a recency effect. Laypeople over-weighed both incriminating and ambiguous evidence, also suggesting a recency effect. These recency effects in updating guilt judgments only occurred when I primed police and laypeople with efficient social norms and not with thorough social norms. Finally, there were mixed results between samples regarding evaluations of ambiguous evidence. Police participants evaluated ambiguous evidence as more incriminating when they encountered incriminating evidence before ambiguous evidence, but only in the efficiency norms condition and not in the thoroughness norms condition. Laypeople participants evaluated ambiguous evidence as more incriminating in the thoroughness norms condition than in the efficient norms condition.

Primary Conclusions

A common thread throughout all three papers is using LRs to communicate evidence strength. Paper #1 discussed the benefits of using LRs and log LRs to communicate evidence strength, Paper #2 found police more accurately understood DNA evidence when presented in an LR format, and Paper #3 used LRs to manipulate the incriminating evidence. Of interest is that I found recency effects in Paper #3, despite using LRs to communicate evidence strength. Even though I found police interpreted LRs more accurately when judging suspect guilt when there is one piece of evidence in Paper #2, perhaps this effect changes when there are multiple pieces of evidence. Paper #3 did not measure guilt judgments using the same measures as Paper #2, so it is unknown the

extent to which participants were accurate in assessing guilt with the current data I collected in Paper #3, as it is outside the scope for that paper.

Paper #1 provided a framework for understanding evidence strength that can be found in both Paper #2 and Paper #3. For example, findings from Paper #2 support the assertions from Paper #1 that LRs are a good method to communicate objective evidence strength. Paper #1 called for more research comparing subjective to objective evidence strength, which Paper #2 addressed. Paper #3 used the definition of evidence strength provided in Paper #1 when measuring guilt judgments, such that it measured guilt using competing hypotheses of innocence versus guilt by asking participants "*To what extent is the suspect guilty or innocent of committing murder*?" from 1 (*completely innocent*) to 9 (*completely guilty*). Thus, Paper #3 had a goal to measure subjective evidence strength to identify biases in guilt judgments.

Overall, these papers suggest that evidence strength presented in an LR format and social norms of thoroughness are two factors that can improve police investigations. Conversely, evidence strength presented in an RMP format and social norms of efficiency can lead to inaccuracy and bias.

These papers also have implications to improve decision-making within law enforcement. First, Paper #2 has implications for increasing literacy when police encounter statistical information. One method to identify whether police are correctly understanding LRs and RMPs is to show one probability format and examine whether police can translate it into the other probability format, such as identifying whether police can translate information presented as an LR into an RMP and vice versa. Such a method could be implemented into police trainings to facilitate understanding probability

statements regarding evidence strength.

Second, Paper #3 has implications that promote changing norms within a police department in the direction of greater investigative thoroughness. Although changing department norms in the direction of thoroughness is important to decrease biased judgments, in practice this change might be challenging. Police face time pressures with heavy workloads to make investigative decisions (Ask & Granhag, 2005, 2007) and complex investigations, such as murders, require many resources (Fahsing & Ask, 2013). In practice, to what extent can police increase the thoroughness of their investigations within the constraints of limited time and resources? Changing the social norms within a police department could begin with the highest-ranking police officers, as police departments are a very hierarchical organization (Engel & Worden, 2003). For example, officers adopted what they perceived as supervisors' goals, even when these goals did not reflect officers' own goals and attitudes (Engel & Worden, 2003). Thus, changing department norms could begin with higher-ranking officers. Although beyond the scope of changing social norms, some researchers recommend checklists as a method to decrease biased decision-making because they provide a framework to assess decision-making (Marsh, 2009). Thus, checklists could be a tool to promote thoroughness through an investigation.

Limitations and Strengths

Limitations

As with any project, there were several limitations to this dissertation that warrant discussion. This section discusses limitations related to recruiting police officers, convenience samples, experimental realism, focusing on guilt judgments, and the

measures in Paper #2.

Recruiting Police Officers

Paper #1 highlights the importance of more research on police and police investigations, which Paper #2 and Paper #3 attempted to accomplish. However, police officers remained a difficult-to-recruit population. Paper #2 recruited police officers using CloudResearch, but CloudResearch could only recruit 200 police participants total and thus I could not recruit police for both papers. Paper #3 recruited police officers through several methods, including contacting departments and posting the study online. However, these methods resulted in a low response rate.

There are a few possibilities that might have contributed to police being difficult to recruit in Paper #3. Police officers can be difficult to incentivize to participate in research. For example, some departments cannot accept monetary compensation and therefore cannot be incentivized by raffles or payment. Police also might not trust that the studies they complete are anonymous. One police contact mentioned that police are suspicious that researchers track their identities when they complete online surveys.

It is also possible the period of time during which I recruited police contributed to the recruitment difficulties in Paper #3. First, there were police staffing shortages (Calvan & Seewer, 2021), some of which might have been due to COVID-19 being the biggest cause of death among police officers in 2021 (Nickeas & Krishnakumar, 2021). Second, some police were unionizing and sent home after going on strike against COVID-19 vaccine mandates (Calvan & Seewer, 2021). Third, I completed my study only a year after the Black Lives Matter movement became a matter of heightened public interest after George Floyd, a Black man, was killed by a White police officer in May 2020.

Between 15-26 million people participated in protests in reaction to George Floyd's murder in what is perhaps the largest movement in U.S. history (Buchanan et al., 2020).

The crux of the problem, therefore, is that police officers are a hard-to-recruit sample. Because police decisions are, in many cases, the first step in a process that ultimately leads to convicting suspects, it is of critical importance to not only understand police officers' judgments but also to optimize their judgments. This dissertation highlights an urgent need for police involvement in psychological research.

Perhaps researchers can use convenience samples (e.g., college students, MTurk workers) instead of recruiting police officers after a body of research demonstrates these samples make similar decisions to actual police officers. For example, Paper #3 found final guilt judgments were similar between police and MTurk samples. Other research has also found similar results between police and college samples (Charman et al., 2017; Yang et al., 2022). Until more research compares police to other convenient samples, future scholars should aim to identify methods to involve police in their research. Future research can also use within-subjects designs when recruiting police to account for the potential of a low response rate.

Convenience Samples

Relatedly, this dissertation was limited by recruiting convenience samples. Paper #2 recruited police officers through CloudResearch, which integrates with MTurk to recruit participants. People self-select to work on MTurk; thus, this sample could be limited by a self-selection bias. Because CloudResearch could only recruit 200 participants, Paper #3 used a hybrid snowball sample to recruit police. This sample for Paper #3 benefitted from attempting to recruit police officers beyond a participant pool

by directly contacting police departments within my network and from a random sample of all police officers. However, Paper #3 suffered from a low response rate from both the snowball sample and the random sample. Thus, there could be differences between police who self-selected to participate in Paper #2 and Paper #3 compared to police who did not agree to participate.

Experimental Paradigms and Realism

As with any experiment, the paradigms used in Paper #2 and Paper #3 lacked ecological validity to maintain experimental control. For example, participants in Paper #2 read there was a fifty-fifty chance the suspect was guilty to ensure every participant began with the same baseline of guilt regarding the suspect. However, such a baseline might not be realistic. Further, police in Paper #2 simply read a description of one piece of evidence and were asked to judge the suspect's guilt. Even though Paper #3 incorporated more than one piece of evidence, the evidence was still presented in a written format that lacked the realism of actual evidence.

Focus on Guilt Judgments

Although understanding how police form their guilt judgments during an investigation is an important first step, these papers do neglect an obvious future direction: How do these guilt judgments actually affect decisions? There could be statistical differences in guilt judgments depending on certain factors (e.g., evidence type, evidence order), however the present research cannot answer whether these statistical differences meaningfully affect decisions. Thus, a fruitful area of future research is to examine when and how guilt judgments affect police decision-making during an investigation.

The context of a police investigation could lend itself as a situation where judgments do predict decisions because many investigative decisions require standards of guilt. For example, police only need to be merely suspicious to bring a suspect into questioning (Scherr et al., 2020), whereas they need to judge the suspect as more than 50% likely to be guilty to arrest a suspect (Roberts, 2019). Future research can examine whether these standards of guilt match actual police practice.

Measures in Paper #2

The measures used in Paper #2 could be problematic when considering them in the context of the operational definition of evidence strength promoted in Paper #1. Specifically, the log-scale guilt measure used a negatively stated hypothesis for the lower range of the scale below the midpoint. Future research should consider examining a log-scale guilt measure that includes a positively stated alternative hypothesis to align more with the message promoted in Paper #1.

Strengths

Despite the limitations, this dissertation offers strengths through two primary avenues: Intellectual merit and broader impact. Each avenue is discussed in turn.

Intellectual Merit

Despite the difficulties associated with time and resources to recruit police officers in Paper #2 and Paper #3, this dissertation still managed to recruit police officers for both papers. Much psychology and law literature with police implications did not recruit a police sample. For example, there are studies on interrogations (e.g., Hill et al., 2008; Kassin et al., 2003) and evidence order during police investigations (e.g., Dahl et al., 2009; Price & Dahl, 2014) that recruited a student sample instead of a police sample.

The topic of evaluating statistical formats of evidence strength in Paper #2 was the first to be studied among a police population, as previous studies had been conducted on mock jurors (e.g., Martire et al., 2014; Thompson & Newman, 2015). Thus, these dissertations furthered intellectual merit by recruiting from the population of interest.

This dissertation also provided meaningful intellectual merit to enrich the field of psychology and law through an interdisciplinary approach. Specifically, these papers combined statistical principles, cognitive psychology, and social psychology to identify sources of police bias and propose a pathway to reduce bias. Paper #1 used statistical principles to clarify definitions of evidence strength and to distinguish between subjective and objective evidence strength. Paper #2 used cognitive psychology based on the theory that people understand frequencies better than ratios (Gigerenzer & Hoffrage,1995) to test police understanding of evidence strength formats, surprisingly finding the opposite: Police understood the evidence more accurately in a ratio format than in a frequency format. Paper #3 combined cognitive psychology and social psychology by examining how order (cognitive) and norms (social) affected judgments of suspect guilt, finding social norms of thoroughness minimized cognitive bias. Thus, these papers addressed psychology and law problems from an interdisciplinary perspective.

Finally, these dissertation papers provide intellectual merit through identifying factors that can contribute to inaccuracy in evaluating evidence to judge suspect guilt (Paper #1, Paper #2), as well as identifying a social psychological mechanism that minimized bias (Paper #3). Together, these papers both identify and propose a method to minimize bias during an investigation, building upon previous psychological research that only identified areas of bias within an investigation (e.g., Charman et al., 2016).

Broader Impacts

Together, these papers also contribute to broader impacts that inform evidence-based practices for a more just society. Considering Paper #1, inconsistent operationalizations of evidence strength could lead to inaccuracy when police use evidence to infer suspect guilt if their decisions are based on research that does not have a unified operationalization of evidence strength. Paper #1 brings this problem to the forefront and proposes using LRs and log-likelihood ratios in practice as methods to convey evidence strength during investigations.

Paper #2 and Paper #3 identify novel factors that affect police investigations. Namely, Paper #2 examined how evidence strength format and evidence type interact to affect judgments of suspect guilt and Paper #3 identified how evidence order and social norms interact to affect judgments of suspect guilt. Based on my findings, police should use LRs when conveying evidence strength and promote norms of investigative thoroughness during investigations to accomplish their goals of accurately incriminating guilty suspects and avoiding wrongfully incriminating innocent suspects.

Results from Paper #2 have implications for ensuring justice after a crime. If police are under-weighing evidence, culprits might not be arrested or convicted. If investigators are too over-weighing evidence, then an innocent suspect could be wrongfully incriminated and convicted. Police departments could incorporate these findings into their police training to ensure proper evaluation of evidence presented in statistical formats.

Paper #3 has implications by identifying that social norms of thoroughness can reduce recency biases. Because social norms of thoroughness reduced bias, police

departments could use these findings in their police training to cultivate a culture that prioritizes thorough investigations, rather than a culture that prioritizes efficient investigations that could be error-prone. Minimizing biases in police investigations would ultimately ensure the correct perpetrator is incriminated and convicted to reduce wrongful convictions.

This dissertation also contributes to broader impacts through disseminating findings. Results from Paper #3 were presented at the American Psychology-Law Society annual conference in 2022. This conference was attended by psychology and law professionals who can build upon this work.

Future Research

These papers provide the groundwork for future research in several ways. First, Paper #1 was limited by providing recommendations for police investigations in theory rather than collecting empirical data in practice. However, Paper #2 builds upon Paper #1 by empirically examining how police evaluate evidence in statistical formats. Future research can examine other formats, such as comparing LRs to Log LRs.

Second, Paper #3 proposes social norms of thoroughness can minimize bias, but it is unknown how practical it is for police to use this recommendation. There were also still biased evaluations of ambiguous evidence within the police and laypeople sample despite social norms of thoroughness, albeit in different directions. Although Paper #3 used statistical formats of evidence strength to manipulate incriminating and ambiguous evidence, it is unknown the extent to which numerical formats decrease bias. Numerical formats of evidence strength have decreased bias within a mock juror sample (Martire et al., 2014), but Paper #3 alone cannot identify whether numerical formats of evidence

strength affected order effects. Log LRs, as proposed in Paper #1, could also be a method to potentially decrease order effects if they do lead to more accurate interpretations of evidence.

As mentioned in the previous paragraph, promoting social norms of thoroughness could be practically implausible. Third, future research could examine the effectiveness of concrete methods to implement this recommendation, such as a program evaluation study that examines the effectiveness of a police training on promoting norms of investigative thoroughness. Police are subject to time pressure during investigations (Ask & Granhag, 2005, 2007), thus future research should examine ways to promote cognitive thoroughness while expediting investigative procedures.

Fourth, future research should consider examining the influence of gender during police investigations. The police samples in Paper #2 and Paper #3 were mostly men, which tends to correspond with the demographics of most police departments (Starheim, 2019). One difference between the samples in Paper #3, aside from one sample consisting of police (Study 1) and one sample consisting of laypeople (Study 2), is the gender composition of the samples. It is possible that the difference in findings between samples in Paper #3 could be related to gender, the gender ratio was majority men in the police sample (83% men, 10% women) but the ratio was distributed more evenly in the MTurk sample (57% men, 42% women).

Conclusion

Christopher Tapp was wrongfully convicted after police inaccurately evaluated evidence. In other words, police investigators who failed to properly evaluate evidence led them to incorrectly judge Tapp's guilt. These judgments of suspect guilt precede

consequential investigative decisions, such as the decision to arrest. Innocent people entering this loop of decision-making inherently impedes justice, particularly because police are humans prone to bias.

Despite the importance of how police perceive evidence and police's role at the forefront of the criminal justice system, most psychological research focuses on laypeople's evaluations of evidence and evidence strength during a trial (Jang, 2021). This dissertation contained three papers that separately, yet cohesively, furthered our understanding of how police use evidence to judge suspect guilt to fill a gap in previous psychological literature. In sum, this dissertation can contribute to understanding how evidence can affect police judgments of suspect guilt during an interrogation.

References

Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of Investigative Psychology and Offender Profiling*, *2*(1), 43–63. https://doi.org/10.1002/jip.19

Ask, K., & Granhag, P. A. (2007). Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology*, *37*(3), 561–591. https://doi.org/10.1111/j.1559-1816.2007.00175.x

Ask, K., Granhag, P. A., & Rebelius, A. (2011). Investigators under influence: How social norms activate goal-directed processing of criminal evidence. *Applied Cognitive Psychology*, *25*(4), 548–553. https://doi.org/10.1002/acp.1724

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*, 161–164. https://doi:10.1016/j.scijus.2009.07.004

Blair, J. P., & Rossmo, D. K. (2010). Evidence in context: Bayes' theorem and investigations. *Police Quarterly*, *13*(2), 123–135. https://doi.org/10.1177/1098611110365686

Buchanan, L., Bui, Q., & Patel, J. K. (2020, July 3). Black Lives Matter may be the largest movement in U.S. History. *The New York Times*. https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html

Calvan, B. C., & Seewer, J. (2021, October 15). Cities, police unions clash as vaccine mandates take effect. *AP News.* Retrieved, https://apnews.com/article/coronavirus-pandemic-health-police-lawsuits-

d072248b2dd77859373744c696e5cfa7

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, *7*(2), 91–103. https://doi.org/10.1037/1076-898X.7.2.91

Charman, S. D., Kavetski, M., & Mueller, D. H. (2017). Cognitive bias in the legal system: Police officers evaluate ambiguous evidence in a belief-consistent manner. *Journal of Applied Research in Memory and Cognition*, *6*(2), 193–202. https://doi.org/10.1016/j.jarmac.2017.02.001

Dahl, L. C., Brimacombe, C. A. E., & Lindsay, D. S. (2009). Investigating investigators: How presentation order influences participant–investigators' interpretations of eyewitness identification and alibi evidence. *Law and Human Behavior*, *33*(5), 368–380. https://doi.org/10.1007/s10979-008-9151-y

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74–78. https://doi.org/10.1016/j.forsciint.2005.10.017

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119. https://doi.org/10.1038/s41558-018-0360-1

Fahsing, I., & Ask, K. (2013). Decision making and decisional tipping points in homicide investigations: An interview study of British and Norwegian detectives. *Journal of Investigative Psychology and Offender Profiling*, *10*(2), 155–165. https://doi.org/10.1002/jip.1384

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of

arguments. *Journal of Personality and Social Psychology*, *71*(1), 5–24. https://doi.org/10.1037/0022-3514.71.1.5

Engel, R. S., & Worden, R. E. (2003). Police officers' attitudes, behavior, and supervisory influences: An analysis of problem solving. *Criminology*, *41*(1), 131–166. https://doi.org/10.1111/j.1745-9125.2003.tb00984.x

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology*, *13*(2), 357–371. https://doi.org/10.1348/135532507X238682

Hogg, M. A. (2010). Influence and leadership. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470561119.socpsy002031

Horgan, A. J., Russano, M. B., Meissner, C. A., & Evans, J. R. (2012). Minimization and maximization techniques: Assessing the perceived consequences of confessing and confession diagnosticity. *Psychology, Crime & Law*, *18*(1), 65–78. https://doi.org/10.1080/1068316X.2011.561801

Jang, M. (2021). *Impacts of evidence on decision-making in police investigation* (Doctoral dissertation). Retrieved from Electronic Theses Online Service (EThOS). (Order No. uk.bl.ethos.837277)

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

Memory and Cognition, *2*(1), 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, *27*(2), 187–203. https://doi.org/10.1023/A:1022599230598

Lidén, M., Gräns, M., & Juslin, P. (2019). 'Guilty, no doubt': Detention provoking confirmation bias in judges' guilt assessments and debiasing techniques. *Psychology, Crime & Law*, *25*(3), 219–247. https://doi.org/10.1080/1068316X.2018.1511790

Lieberman, J. D., Carrell, C. A., Miethe, T. D., & Krauss, D. A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law*, *14*(1), 27–62. https://doi.org/10.1037/1076-8971.14.1.27

Marsh, S. (2009). The lens of implicit bias. *Juvenile and Family Justice Today*, *18*, 16 – 19. Retrieved, https://www.ojp.gov/ncjrs/virtual-library/abstracts/lens-implicit-bias

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, *240*, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

Murphy, H. (2019, July 18). The jury said he killed her daughter. She helped clear his name. *The New York Times.* Retrieved, https://www.nytimes.com/2019/07/18/us/angie-dodge-christopher-tapp.html

Nickeas, P. & Krishnakumar, P. (2021, October 22). Many police unions are pushing

back on vaccine mandates. Here's why. *CNN.* Retrieved, https://www.cnn.com/2021/10/21/us/police-unions-vaccine-workers-rights/index.html

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

Otterbourg, K. (2021, February 10). Christopher Tapp. *The National Registry of Exonerees.* Retrieved, https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=5592

Price, H. L., & Dahl, L. C. (2014). Order and strength matter for evaluation of alibi and eyewitness evidence. *Applied Cognitive Psychology*, *28*(2), 143–150. https://doi.org/10.1002/acp.2983

Roberts, A. (2018). Arrests as Guilt. *Alabama Law Review*, *70*(4), 987–1030.

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons.

Scherr, K. C., Redlich, A. D., & Kassin, S. M. (2020). Cumulative disadvantage: A psychological framework for understanding how innocence can lead to confession, wrongful conviction, and beyond. *Perspectives on Psychological Science*, *15*(2), 353–383. https://doi.org/10.1177/1745691619896608

Sommers, S. R., & Douglass, A. B. (2007). Context matters: Alibi strength varies according to evaluator perspective. *Legal and Criminological Psychology*, *12*(1), 41–54. https://doi.org/10.1348/135532506X114301

Starheim, R. P., (2017). *Women in policing: Breaking barriers and blazing a path* (Report No. 252963). National Institute of Justice. https://www.ojp.gov/pdffiles1/nij/252963.pdf

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*(1), 99–139. https://doi.org/10.1037/a0021650

Thompson, W. C., & Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, *39*(4), 332–349. https://doi.org/10.1037/lhb0000134

Thompson, W. C., Grady, R. H., Lai, E., & Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, *17*(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*(2), 45–75. https://doi.org/10.1111/j.1529-1006.2006.00027.x

Yang, Y., Madon, S., Cabell, J. J., Moody, S. A., Guyll, M. (2022). *The effect of presumption of guilt on police guilt judgments*. Unpublished Manuscript.

Appendix

Paper #3: Perceptions of Suspect and Victim Race (used to determine whether participants paid attention to the case scenario, as there was no mention of race/ethnicity in the case scenario)

1. What race/ethnicity was the suspect, Samuel Scott?
    a. Black or African American
    b. Asian
    c. White or Caucasian
    d. Latina/o
    e. Native American
    f. Indian
    g. Other
    h. Not Sure
2. What race was the victim, Alexis Schafman?
    a. Black or African American
    b. Asian
    c. White or Caucasian
    d. Latina/o
    e. Native American
    f. Indian
    g. Other
    h. Not Sure