Effects of Legal Instructions on Behavioral and Neural Mechanisms of Decision-Making

By

Lauren E. S. Hartsough

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

December 12, 2020

Nashville, Tennessee

Approved:

René Marois, Ph.D.
Jennifer Trueblood, Ph.D.
David Zald, Ph.D.
Jonathan Lane, Ph.D.

ACKNOWLEDGMENTS

I am in the incredibly fortunate position of having a great number of people to thank for their support throughout this process. Thank you!

First, I would like to express my deep gratitude to my advisor Dr. René Marois for giving me this opportunity and for his support and guidance throughout this research. His innovative thinking inspires my research and his kindness made it possible for me to persevere in this effort.

I would also like to extend my sincere thanks to each of my committee members, Dr. Jennifer Trueblood, Dr. David Zald, and Dr. Jonathan Lane. Each have contributed unique and invaluable insight and discussion that has shaped both this work and my growth as a scientist.

Thank you to all members of the Marois lab, past and present, who have provided support and friendship over the years. I'd also like to acknowledge those who contributed to the present research: thank you to Dr. Matthew Ginther for his patience and guidance in my early years of the program and for his critical contributions as co-author for several of these chapters. Thank you to Nicole Seedarnee for her work collecting behavioral data, and to Leah Mann for her work assisting in the development of scenarios and data collection.

I would also like to thank all of the faculty and staff of the VU Psychology Department for their years of support and encouragement. This process would not have been possible without the dedication to helping students in their research journey that permeates this department.

Last but not least, I want to give my love and thanks to my friends and family.

To Rhynn for his unwavering friendship and support, and for helping me to find magic even in the mundane. To Clair for her friendship and commiseration, who kept me believing that we can do this. To Liv for her years of friendship, and for sharing her humor even in challenging times. To Kirby and Charlie for their antics and companionship.

Thank you so much to Molly, Bryan, and Margaret for their love and support over these years.

Thank you to My Mom and Dad for their Love, support, and sacrifices throughout my life that made this possible. They have been there for me through thick and thin with encouragement and humor, and have made me who I am today.

Thank you to my beloved husband John who has been by my side for every step of this process, with patience and unconditional love and support. He has been my rock and brings joy and laughter into my life each and every day.

TABLE OF CONTENTS

Chapter

LIST OF TABLES

Chapter Three

Chapter Six

LIST OF FIGURES

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

CHAPTER 1

GENERAL INTRODUCTION

We engage in decision-making constantly throughout our daily lives. Understanding the factors that influence decisions is of substantial interest not only to psychological researchers but also to fields that have high-stakes decisions such as marketing, medicine, and law. Indeed, understanding these factors allows for the manipulation of decisions, for instance to motivate consumer purchases or to dictate the application of laws in accordance with the U.S. justice system. Perhaps the most straight forward manner to influence decision making is to instruct the decision maker about how they should make their decision. Psychological research has found that instructions affect decision-making in a number of ways; for instance, the speed-accuracy trade-off shows that individuals' performance can vary depending on whether they are instructed to emphasize deciding as fast as possible or making as few mistakes as possible (e.g. Spieser et al., 2017). Framing effects, in contrast, show that explicit instructions are not required to alter behavior, as the manner in which information is presented may suffice to affect decisions (e.g. Johnson, 1987). The effect of instructions on decision-making in everyday life is especially salient in the U.S. legal system, as individuals are instructed to make decisions according to prescribed legal standards rather than relying solely on their own intuition and experience. The U.S. legal system relies on an adversarial approach to present the facts of a case, as jurors hear competing evidence from each side and must deliberate to determine the facts of the case. For criminal trials, in which an individual is accused of breaking the law, the prosecution brings charges against a defendant. Civil trials, on the other hand, typically involve financial disputes between a plaintiff and a defendant. In both of those cases, two key classes of legal instructions are commonly issued

to jurors to dictate how they should assess evidence when making their decision. First are burdens of proof, which are standards that impose specific decision thresholds that must be met in order to decide in favor of the prosecution/plaintiff; these include *beyond a reasonable doubt* for criminal trials and *by a preponderance of the evidence* for civil trials. The second class are Instructions to disregard evidence, whereby judges may instruct a jury to disregard evidence that has been presented to them if it is deemed inadmissible. In such case, jurors are not to consider the inadmissible evidence when deciding whether or not the burden of proof has been met.

The goals of these legal instructions are to provide impartial and equitable decisions across cases and to reduce the bias of evidence considered irrelevant or prejudicial to the case. However, as discussed below, there is ample evidence that both burdens of proof and disregard instructions are not interpreted or applied correctly by jurors (i.e. in a manner consistent with the goals of the justice system). The US has leading rates of incarceration and harsh sentencing, with defendants in criminal cases facing the possibility of monetary penalty, loss of liberty through incarceration, or even the death penalty (Carlsmith, Darley, & Robinson, 2002; Kyckelhahn, 2015). Legal decisions are costly for society as well as the individuals directly involved, as the average annual cost to incarcerate an individual exceeds $37,000 (Bureau of Prisons, 2019) and the total annual cost of the criminal justice system (incarceration and court proceedings) is over $182 billon (Wagner & Rabuy, 2017). Civil cases likewise have a high cost for individuals and society, as the annual economic impact of tort litigation (e.g. lawsuits brought for personal loss or harm) has been estimated at $263 billion (Towers Watson, 2012). Given these high costs of legal decisions, there is a strong impetus to ensure that these decisions are rendered as equitably as possible. Research on the effect of instructions on legal decision-making not only enhances our psychological understanding of the processes involved in the evaluation of information to drive decisions, but

also has the potential to inform efforts to reform the U.S. justice system. The goal of this dissertation is to use both behavioral and neurobiological approaches to better understand how these two classes of instructions affect legal decision-making.

Burden of Proof Instructions

Jurors are commonly tasked with applying two distinct burdens of proof. For criminal trials the prosecution has the burden of proving beyond a reasonable doubt that the defendant is guilty of the crime they are charged with, while in civil trials the burden of proof is on the plaintiff to prove by a preponderance of the evidence that they are owed damages. Jurors are instructed to find in favor of the prosecution/plaintiff only if the evidence presented satisfied the prescribed legal burden of proof. A preponderance of the evidence is a lenient standard that corresponds to a tipping of the scales such that one side's evidence is more likely true than not, which aligns with a theoretical decision threshold of 50%. Beyond a reasonable doubt (BaRD) is a more stringent burden of proof that favors false acquittals over false convictions of an innocent person (*In re Winship,* 1970). It is typically defined as there being firmly convincing evidence, and is conceptualized as a decision threshold of 90% (ex. Laudan, 2003; Newman, 1993). This standard has its roots in "Blackstone's Ratio", a principle put forth by William Blackstone that it "is better that ten guilty persons escape than that one innocent suffer" (as cited in Newman, 1993). Jurors are told which burden of proof they are required to apply and may receive specific instructions defining the burden of proof. Judges specifically do not permit quantifying these standards for jurors and instead argue that jurors should interpret them as they will, treating legal standards as guidelines for their decision-making.

Studies have consistently demonstrated that laypeople do not assign these burden of proof instructions in a manner that is consistent with the prescribed threshold. Estimates of the BaRD standard for laypeople range between 65% and 90% across studies, with most falling well below the theoretical 90% threshold (Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; Simon & Mahan, 1971). In contrast, individuals tend to be overly stringent in their application of the PoE standard, though few studies have assessed the interpretation of PoE. Simon and Mahan (1971) found that potential jurors' estimates of PoE fell at around 75%, similar to their BaRD estimates, indicating that they do not distinguish between the civil and criminal burdens of proof as they are expected to. Judges are typically more consistent with the prescribed legal standards, with mean estimates for BaRD that were close to 90% and estimates of PoE between 50-55% (McCauliff, 1982; Simon & Mahan, 1971). It is possible that the discrepancy arises from jurors favoring their own innate decision thresholds over the provided legal standards. Indeed, deciding to convict can lower individuals' estimates of the BaRD standard compared to those who acquit, indicating that people may adjust their interpretations of the burdens of proof to align with their own decision threshold (Park et al., 2016). There is some evidence that people have similarly strong aversions to both false convictions and false acquittals, which could also account for discrepancies in how laypeople interpret legal standards. (Arkes & Mellers, 2002).

From this brief review of the literature, it is clear that jurors are often applying these burdens of proof in a manner that is not consistent with the prescribed decision thresholds or with judges' understanding of these instructions. However, no studies have yet sought to determine how individuals' interpretation of these burdens of proof compare to their own intuitive decision thresholds. Given that jurors' estimates for the BaRD standard are typically less stringent than both judges' and the 90% standard (Dhami et al., 2015; Horowitz & Kirkpatrick, 1996; McCauliff,

1982; Simon & Mahan, 1971), while their estimates of the PoE standard are more stringent than the 50% standard (Simon & Mahan, 1971), it is possible that laypersons' intuitive decision thresholds may fall between BaRD and PoE, and that they adjust the PoE threshold upwards and the BaRD threshold downwards in order to have both more closely align with their anchoring intuitive threshold. Furthermore, previous research has not examined how these burdens of proof are applied in other domains of society. Arbitrary decision thresholds are not unique to the legal field, as they are found in other spheres of society with high cost decisions. Doctors, for example, must determine whether the benefit-to-risk ratio to their patients meets the threshold for treatment, and scientists apply standardized statistical thresholds to report advancements in the field and assess replicability. A universal question across these domains is what constitutes enough evidence to reach a decision, and comparing standards across social domains may shed light on the common elements of decision thresholds.

Disregard Instructions

Judges may issue a disregard instruction to jurors when they have heard evidence that has been ruled inadmissible for a number of reasons, either because it was obtained illegally, is hearsay, is not relevant to the case, or is prejudicial (Federal Rules of Evidence). Typically, the admissibility of evidence is determined in hearings in advance of the trial and not in front of a jury. However, it is not uncommon during testimony for a lawyer or witness to disclose evidence that has not been ruled on or has been ruled inadmissible. This requires the opposing counsel to object and the judge to make a ruling about this evidence. If the evidence is considered inadmissible, it is stricken from the record and the jury is instructed to disregard it. The disregard instruction is typically repeated as part of the jury instructions provided immediately prior to deliberation.

Character evidence is one common form of evidence that is considered inadmissible in criminal trials (Rule 404 of the Federal Rules of Evidence). Specifically, the prosecution cannot present evidence related to the defendant's character that shows the defendant's propensity to commit a crime including any prior "crimes, wrongs, or other acts". Previous work has suggested that when jurors hear evidence of prior convictions, they are more likely to convict compared to when they hear of a prior acquittal or no priors (Greene & Dodge, 1995), and that defendants viewed as having negative character traits are perceived as guiltier and receive greater suggested sentences (Izzett & Leginski, 1974; Kaplan & Kemmerick, 1974; Landy & Aronson, 1969; Wissler & Saks, 1985).

Prior research on whether a disregard instruction is effective in reducing the bias has not come to a general consensus. While some findings suggest that jurors can follow directions to disregard certain types of inadmissible evidence (Simon, 1966) or even overcompensate in correcting for the evidence (Thompson et al., 1981), others point to jurors' inability to ignore evidence to the extent that is equivalent to never having heard it in the first place (Carretta & Moreland, 1983; Casper et al., 1989; Edwards & Bryan, 1997; Lieberman & Arndt, 2000; Sue, Smith, & Caldwell, 1973; Tanford & Cox, 1988). Individuals are more likely to correctly apply disregard instructions when they understand and agree with the reason for the inadmissible ruling, for hearsay evidence, or when the evidence favors acquittal of the defendant rather than conviction (Kassin & Sommers, 1997; Thompson et al., 1981). Yet other studies suggest that the disregard instruction can have the opposite effect of what is intended, as strong admonishments to disregard inadmissible evidence or provide legal reasoning can lead to higher rates of conviction and higher damages against the defendant (Broeder, 1959; Pickel, 1995; Thompson et al., 1981). While some have interpreted this back-firing effect as reactance by jurors who feel that the judge is attempting to influence their decision (Lieberman & Arndt, 2000; Thompson et al., 1981), others suggest

instead that the disregard instruction inadvertently draws attention to the inadmissible evidence and emphasizes to jurors that it is important in some way, leading to the back-firing effect (Broeder, 1959; Lieberman &Arndt, 2000; Pickel, 1995). Further, attempts at thought suppression may simply make the information more salient and accessible (Butler & James, 2010; Edwards & Bryan, 1997; Lieberman & Arndt, 2000). Hindsight bias may also play a role in preventing individuals from being able to fully disregard evidence, as they have integrated it into their interpretation of the facts, thus making it difficult to ignore (Casper et al., 1989). This interpretation is consistent with the Story Model of juror decision-making, which posits that jurors cognitively represent evidence in the form of stories that forms a narrative that can be compared to potential verdicts (e.g. Pennington & Hastie, 1986; Pennington, 1993). Jurors may simply fail to recognize that their narrative has incorporated the inadmissible evidence. In fact, judges show the same inability to disregard relevant but inadmissible character evidence (Wistrich, Guthrie, & Rachlinski, 2005). It remains unclear, however, how jurors attempt to suppress information when given a disregard instruction and how these instructions may alter punishment decision processes. Previous studies have not incorporated neuroimaging data into understanding how inadmissible character evidence and disregard instructions affect punishment, which could shed additional light on the mechanisms involved.

Present Research

The research presented in this thesis assesses the behavioral and neural effects of legal instructions on legal decision-making and compares these decisions to both intuitive decision thresholds as well as those employed in other societal domains. Chapters 2-4 implement a psychometric approach to evaluate the effect of decision burden of proof instructions (i.e. PoE, BaRD) on

decision-making. The use of psychometric functions to assess more complex decisions is a distinguishing feature of this research, as this approach has typically been applied in psychophysical studies. Chapter 2 evaluates the effect of PoE and BaRD instructions on laypeople's culpability decisions to determine whether they have the intended effects on decision-making, and to determine how such burdens of proof compare to people's intuitive sense of culpability (i.e. in the absence of any legal instruction). We also test in this Chapter the generality of the application of these instructions in the legal, non-legal, and psychophysical domains. Our results indicate that individuals are more stringent for the BaRD versus PoE instruction, but that they are also more stringent for PoE versus the intuitive belief (IB)/non-instructed condition and versus the legally prescribed threshold of 50%, a finding that was consistent across domains. These findings suggest that the overly stringent application of PoE is not due to individuals interpreting the standard to align it with their intuitive belief. We also found that participants were also more stringent in the legal versus non-legal domain, possibly because of the potentially serious consequences a legal decision can have to defendants. Chapter 3 further examines the generality of the application of the legal instructions across legal and non-legal domains (medical, scientific, misc.), and assesses the effect of expertise on these decisions among different expert groups. We found that legal experts applied the PoE and BaRD instructions as intended regardless of domain and also used a decision threshold consistent with PoE in the absence of any instruction. Our other expert groups (medical, scientific, humanities) were generally more conservative in their decisions, including their interpretation of PoE, and showed more conservative decisions in the legal domain. As in Chapter 2, our non-expert 'control' participants were more stringent in their decision in the legal than in non-legal domains, and this effect was especially pronounced for PoE. Chapter 4 aimed at understanding why PoE is interpreted more stringently than the legally

prescribed standard by laypeople, the very individuals likely to serve as jurors. Specifically, we compared multiple instructions that were equivalent to PoE across domains and in comparison to intuitive decision. We found that the phrase "preponderance of the evidence" itself contributes to more stringent decisions, but only within the legal domain, thus suggesting that it's the interaction of a legal context and the PoE instruction that lead to the overly stringent application of this standard. Chapter 5 aimed at understanding why legal domain scenarios are consistently adjudicated more conservatively than any other domain scenarios. Specifically, we hypothesized that it is the inherent costs (especially to the defendant) to rendering a legal decision that makes these high-staked decisions conservative. We assessed how the cost of the decision influences decisions across domains and instructions. We found – surprisingly - that the absolute decision cost had little influence on laypeople's decision thresholds across domains, and that legal decisions were more conservative regardless of the cost. Together, these studies indicate that legal instructions have complex effects on decision-making that are dependent on the context in which the decision takes place and on the decision-maker's expertise in the relevant domain. Finally, Chapter 6 evaluated the effect of instructions regarding the admissibility of character evidence on third-party punishment (TPP) behavior and on the brain mechanisms supporting TPP. Specifically, we first assessed the effect of instructions on the admissibility of character evidence (i.e. 'must disregard', 'may consider') on punishment decisions to determine whether they have the intended effects on decision-making, and to determine how these decisions compare to people's intuitive punishment decisions (i.e. in the absence of any legal instruction). We then used fMRI to examine how both character evidence and admissibility instructions affect the TPP neural network by adapting Ginther et al.'s (2016) paradigm from the Marois lab. We found that the admissibility

instruction influenced punishment decisions even in the absence of an effect of character evidence, both behaviorally as well as neurally. I conclude this thesis in Chapter 7 by a General Discussion.

CHAPTER 2

EFFECT OF LEGAL STANDARDS AND SOCIETAL DOMAINS

Introduction

Decision-making – something we all engage in 100 times a day – is of substantial interest not only to psychological researchers but also to fields with high-stakes decisions such as marketing, medicine, and law. A fundamental question in understanding decisions is what constitutes enough evidence to pass the threshold needed to make a choice between different options (for instance- choosing product A over product B, diagnosing a patient with a specific disease, finding a defendant guilty). In particular, the U. S. legal system provides a compelling real-world context highlighting the importance of decision thresholds, as jurors are instructed to apply prescribed legal thresholds – that is, burdens of proof – rather than relying on their own intuition and experience.

There is a strong impetus for ensuring that legal decisions are well-rendered given the prevalence and high cost of such decisions for both the individual and society. Defendants in criminal cases face the possibility of monetary penalty, loss of liberty through incarceration, or even the death penalty. The world-leading rates of incarceration and harsh sentencing in the US consumes vast amounts of resources (Carlsmith, Darley, & Robinson, 2002; Kyckelhahn, 2015), with the average annual cost of incarceration exceeding $37,000 (Bureau of Prisons, 2019) and the total annual cost of criminal court proceedings and incarceration surpassing $182 billion (Wagner & Rabuy, 2017). Civil cases, in which plaintiffs seek monetary compensation from a defendant, also have astronomical costs; the annual economic impact of tort litigation (ex. lawsuits brought for personal loss or harm) has been estimated at $263 billion (Towers Watson, 2012), while the

total cost of class action litigation (ex. a lawsuit brought by a group of people suffering the same harm) has exceeded $939 billion in the US in 2018 (Boettrich & Starykh, 2019). Yet, little is known about the internal decision processes individuals may engage in the course of rendering these high-cost decisions. Jurors may be issued instructions but these are purposefully vague and may not even influence decisions in the manner intended by the legal system (see below). Therefore, understanding how individuals interpret legal instructions to reach a decision has the potential to inform efforts to reform the U.S. justice system. This chapter not only examines how instructions to apply civil and criminal decision standards influence decisions, but also compares these standards to individuals' intuitive decisions thresholds across several social and perceptual domains in order to provide insight into the utilitarian properties of legal standards.

Legal Burdens of Proof

The U.S. legal system actively imposes burdens of proof (decision thresholds) for judges and jurors to follow, with the goal of providing equitable decision thresholds across cases. The adversarial nature of our justice system means that judges and juries typically hear competing versions of the evidence, and must apply established legal criteria to determine whether or not the plaintiff/prosecutor has met the prescribed burden of proof. Civil cases, in which a plaintiff seeks financial compensation, use a **preponderance of the evidence (PoE)** as the burden of proof to determine the liability of a defendant. Criminal cases on the other hand require the more stringent proof of **beyond a reasonable doubt (BaRD**), which favors releasing a guilty offender over falsely convicting an innocent person (*In re Winship*, 1970). Most legal scholars consider a PoE to be a tipping of the scales such that one side's evidence is more likely than not. In principle, this should align with a decision threshold of just over 50%. In contrast, BaRD is typically defined as

there being firmly convincing evidence, with a theoretical decision threshold of 90% (ex. Laudan, 2003; Newman, 1993). This threshold has its roots in "Blackstone's Ratio", a principle put forth by William Blackstone that it "is better that ten guilty persons escape than that one innocent suffers" (as cited in Newman, 1993). Jurors typically are told which burden of proof they are required to apply in their decisions and may or may not receive additional definitions describing the burden of proof. Judges specifically do not permit quantifying these standards for jurors and instead argue that jurors should interpret them as they will, treating legal criteria as guidelines for subjective decision-making.

Studies have assessed the extent to which jurors' interpretations of the instructions align with the ideal thresholds for these burdens of proof. When asked to assign probabilities to the BaRD burden of proof, potential jurors' estimates range between 0.65 and 0.90 across studies, with most falling well below the theoretical threshold (Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; Simon & Mahan, 1971). Judges' mean estimates for BaRD were close to 90%, though there was still variation in their responses (McCauliff, 1982; Simon & Mahan, 1971). Teitcher and Scurich (2017) used logistic regression to compare implicit decision thresholds for the BaRD burden of proof rather than having participants provide an estimate explicitly. These thresholds fell at or below 80% across several levels of crime types and punishment outcomes. In contrast to BaRD, few studies have assessed how laypeople interpret the PoE threshold; Simon and Mahan (1971) found that potential jurors' estimates of PoE were around 75% indicating that they distinguish less between civil and criminal burdens of proof. Judges place the PoE standard between 50-55% (McCauliff, 1982; Simon & Mahan, 1971). It is possible that these discrepancies in the interpretation of the burdens of proof between judges and jurors is due to jurors favoring their own decision thresholds over the provided standards. This hypothesis is

consistent with the finding that individuals who decide to convict tend to ascribe lower estimates of BaRD compared to those who acquit, indicating that people may adjust interpretations of the burdens of proof to align with their own decision thresholds (Park et al., 2016). One factor driving this difference between jurors' decisions and the prescribed legal standards may be that people have similarly strong aversions to both false convictions and false acquittals, in contrast to the legal standard of BaRD which favors false acquittals to false convictions at 10:1 (Arkes & Mellers, 2002).

The above studies illustrate the extent to which jurors' decision-making standards may be mis-aligned not only to those of judges, but also to the prescribed burdens of proof. Given this state of affairs, it is all the more surprising that no studies have yet sought to determine how the burdens of proof thresholds compare to individuals' intuitive decision thresholds. Given that jurors' estimates for BaRD are typically less stringent than both judges' and the theoretical standard (Dhami et al., 2015; Horowitz & Kirkpatrick, 1996; McCauliff, 1982; Simon & Mahan, 1971), while their estimates for PoE are more stringent than the intended construct (Simon & Mahan, 1971), it is conceivable that laypersons' intuitive decision thresholds fall between BaRD and PoE. According to this scheme, jurors may adjust the PoE threshold upwards and the BaRD threshold downwards in order to have both align more closely with their anchoring intuitive threshold.

Not only do we not know how legal burden of proof thresholds compare to laypeople's intuitive heuristics, we have very little idea of how generalizable the application of such instructions is beyond the legal context. How do legal standards compare to decision thresholds in other societal domains, and to what extent are the effects of these decision standards specific to the legal context? These questions can be answered by comparing the effects of different burdens

of proof in both legal and non-legal (e.g. scientific, medical, financial) contexts. They can even be extended from societal domains to the physical domain by assessing the extent to which these decision thresholds affect our judgement of the perceptual world.

Psychometric Approach

Comparing the effects of different burden of proof instructions across a wide range of domains requires bringing them into a common experimental space. To do so, here we have adopted a psychometric approach widely applied in psychophysics. Psychometric functions are simple mathematical functions that express how changes in a given variable/parameter contribute to changes in decisions (Wichmann & Hill, 2001a). They are frequently used in the field of psychophysics to describe, for instance, the relationship between stimulus contrast and decision response for simple decision processes such as visual or auditory stimulus detection. This analytical approach has not been as widely applied to more complex decisions (ex. legal, medical, etc.). Yet, a psychometric analysis is advantageous due to its simplicity as well as its ability to describe characteristics of the relationship between the dependent and independent variables with just four parameters: one defining the functions' position on the abscissa (often referred to as the threshold), one defining its slope, and one each defining the upper and lower as asymptotes (Klein, 2001; Kroll et al., 2002). To compare how different variables may affect the decision-making threshold it is customary to use the 50% mark, which represents the point at which the stimulus strength along the abscissa is sufficient to change the decision (from no to yes for example). The slope of the psychometric function reflects the rate of change of the decision with respect to the stimulus strength and serves as a measure of the strength of the relationship between the variable

and decision, while the upper and lower asymptotes convey the likelihood of making a decision when the stimulus strength is at its maximum and minimum values, respectively.

The rationale for applying a psychometric approach to legal decision-making is multi-fold: it not only provides a quantitative and sensitive assessment of the effects of manipulating variables on decision-making, but also can reveal specific effects of these variables onto distinct parameters of the psychometric function, thus potentially uncovering not just whether but also how different variables affect legal decision-making. Furthermore, while previous studies of legal decision-making have typically manipulated the evidence strength dichotomously as either weak or strong (e.g. Horowitz & Kirkpatrick, 1996; Kagehiro & Stanton, 1985; Park et al., 2016), a psychometric approach lends itself to examining decisions across a range of evidence strengths that include moderate, ambiguous strength levels. This may be more applicable to legal decisions as research suggests that cases in which the evidence against a defendant is clearly weak or clearly strong are more likely to be dropped or settled via a plea bargain, while "close" cases are more likely to make it to a jury trial, the very situation where these burdens of proof are applied by laypeople (Champion, 1989; Lederman, 1999).

Present Study

Here we implement an exploratory psychometric approach to assess and compare decision thresholds across legal, non-legal, and psychophysical domains. The main objectives of the study are to compare the effects of the civil (PoE) and criminal (BaRD) burdens of proof on decisions relative to subjects' intuitive beliefs (IB, with no legal instruction), and to determine the extent to which the effects of these decision standards are contingent on specific social and physical contexts. To our knowledge, no previous studies have assessed how individuals' application of

legal standards compare to their intuitive decisions or how generalizable the application of these standards is across different domains (legal, non-legal, psychophysical). Furthermore, the psychometric approach should allow us to quantitatively characterize the mechanisms by which instruction and context affect decision-making. Specifically, in Experiment 1 participants rendered a decision to a single text scenario that manipulated the strength of evidence, the decision criteria instruction (IB, PoE, BaRD) and the scenario context (legal, non-legal). Experiments 2 and 3 extended the results of Experiment 1 to the perceptual domain by manipulating the strength of evidence and decision criteria instruction in a classic psychophysical dot motion coherence task.

Experiment 1

Experiment 1 assessed the effect of legal burden of proof instructions on decisions across both legal and non-legal contexts. Participants made a decision in response to a single textual scenario that varied across subjects in the strength of the available evidence, in the decision criteria instruction to be applied (IB, PoE, or BaRD), and in the scenario context (legal or non-legal).

Methods

Participants

We recruited 6067 participants (53% male, Mean age=35.25 years) from the United States via Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013). We eliminated 599 participants for failing attention checks about the content of the trial/instructions (see Supplementary Materials Section 1, Figs S1-S3), yielding a total of 5468 participants included in our analyses. Participants were paid $0.40 for completing the study, which took less than five minutes on average. Past research suggests that 40 responses per group provides strong coverage of bootstrap confidence intervals for psychometric parameters (i.e. intervals more likely to contain

the true value; Fründ, Haenel, & Wichmann, 2011). We therefore recruited participants until we reached at least 40 responses per cell (context x objective evidence strength x instruction) after all exclusion criteria had been applied. All participants provided informed consent, and the experimental protocol was approved by the Vanderbilt University Institutional Review Board.

Design and Materials

Each participant read and responded to a single randomly assigned scenario that was one paragraph in length. The task employed a 2 (scenario context: legal or non-legal) x 7 (objective evidence strength; 20%, 40%, 60%, 80%, 95%, 99%, 100%) x 3 (decision criteria instruction; IB, PoE, BaRD) between-subjects design.

The experiment was administered using the Qualtrics online survey platform. Figure 1 shows a sample trial (Legal context, drug theft x finger print scenario, BaRD instruction), and all of the scenarios used are included in Section 10 of the Supplementary Materials. Participants were first presented with the scenario text on their computer screen. The decision criteria instructions then appeared directly below the scenario on the same screen once participants pressed a button to continue after reading through the scenario. Participants were told that reading the criteria instructions carefully was critical for completion of the study. After reading the instructions, participants responded to a question that appeared on the same screen under both the scenario and instructions. Thus, both the scenario and the instructions were available to participants while making their decision. Evaluation of the scenario/instructions and the subsequent decisions were self-paced.

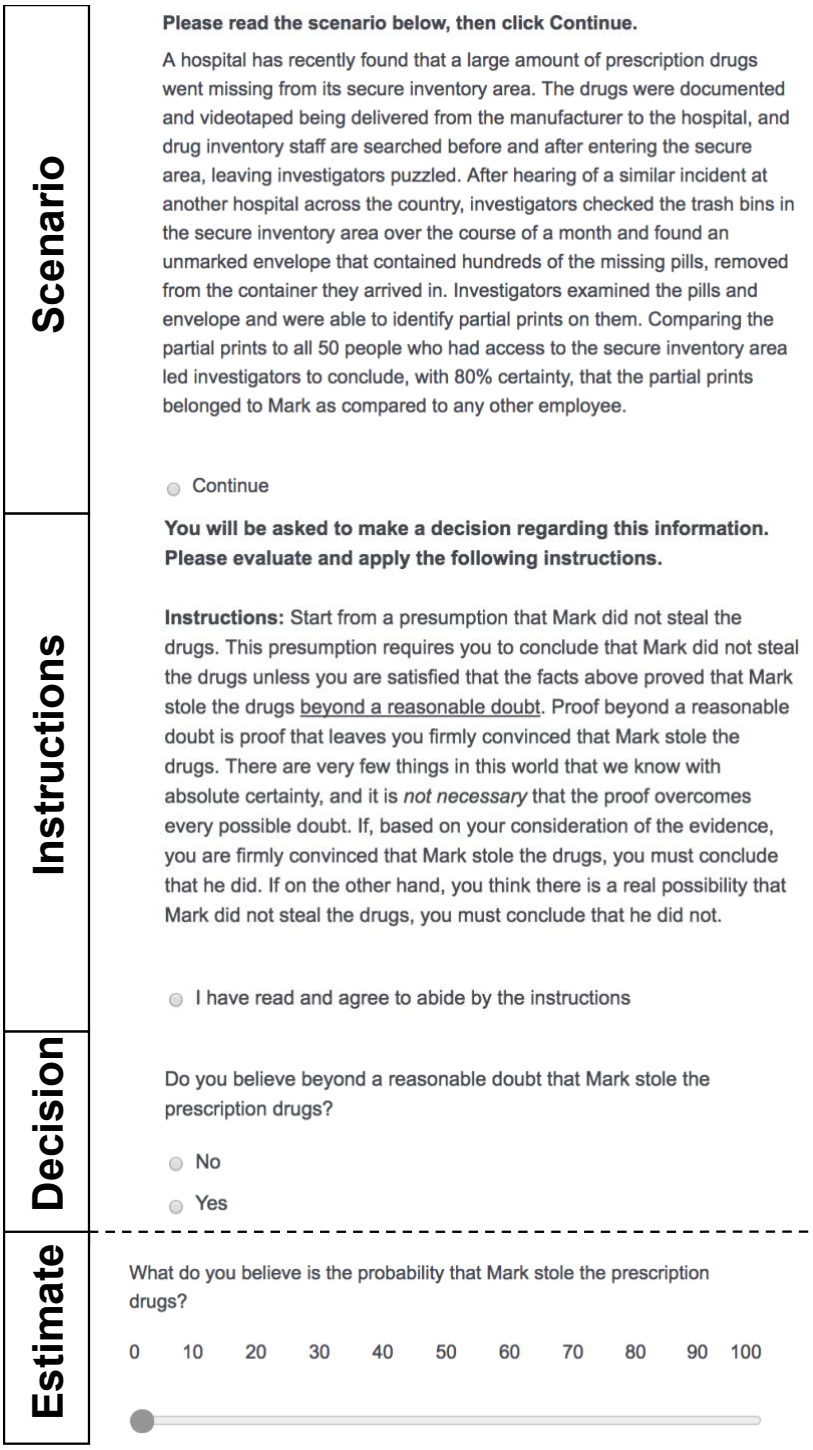

Figure 1. Sample trial as seen by participants in the legal context (prescription drug theft x finger print evidence scenario) with BaRD instruction. Participants read the scenario then pressed "Continue". The instructions then appeared directly below the scenario for the PoE and BaRD conditions; participants clicked "I have read and agree to abide by the instructions" and the decision prompt appeared on the screen. After making their response, participants provided their subjective probability estimate on a new screen.

Participants were randomly assigned to a single scenario that described either a legal or non-legal context and presented the objective evidence. The subject matter of the **legal scenarios** involved a protagonist, Mark, who may have engaged in conduct that is widely accepted as criminal or civil wrong-doing. To cover a broad spectrum of potential legal contexts, the scenarios varied in the fact pattern (stealing prescription drugs, stealing company data, and murder) and in the type of evidence available (video facial recognition, fingerprints, and DNA), thus forming nine possible legal scenarios. The objective evidence strength was presented within the scenario as the level of certainty with which investigators were able to link the available evidence to the protagonist. This evidence strength was communicated with a frequentist measure of probability that varied between subjects across seven possible levels: 20%, 40%, 60%, 80%, 95%, 99%, and 100% (e.g. Investigators concluded with 40% certainty that the DNA found belonged to the protagonist). To increase the realistic interpretation of these probabilities (Wells, 1992; M.G. & R.M, unpublished data), the legal scenarios were crafted so that they described 'closed systems' in which the person who committed the offense was a member of a finite group of individuals (for example, the culprit can only be among the individuals aboard a ship). We compared decisions by fact pattern and by evidence type within the legal context and found that decision thresholds were not significantly different between fact patterns or evidence type (Supplementary Figs S4-S5 and Tables S1-S2); we therefore collapsed the data across fact pattern and evidence type in the results section below.

The subject matter of **the non-legal scenarios** eschewed legal or wrong-doing matters and instead described situations that required participants to render a decision about the occurrence of an event in one of five distinct fact patterns. Specifically, participants were tasked to make a judgement about the likelihood of either: a patient developing Huntington's disease, a stock

underperforming in the market, abnormal water temperatures developing in the Pacific Ocean, the occurrence of a petroleum spill, or the presence of electronic spam information. The objective evidence strength was presented as a frequentist probability that the condition or event occurred or would occur, and varied between 20%, 40%, 60%, 80%, 95%, 99%, and 100%. As in the legal context, we compared decisions by fact pattern within the non-legal context (this context did not have different evidence types), and found that decision thresholds did not differ significantly between fact patterns (Supplementary Fig. S6 and Table S3); we therefore collapsed the data across fact pattern in the results section below.

Below we provide an example of both a legal and non-legal scenario.

> Sample legal scenario: Prescription drug theft x DNA evidence
> A hospital has recently found that a large amount of prescription drugs went missing from its secure inventory area. The drugs were documented and videotaped being delivered from the manufacturer to the hospital, and drug inventory staff are searched before and after entering the secure area, leaving investigators puzzled. After hearing of a similar incident at another hospital across the country, investigators checked the trash bins in the secure inventory area over the course of a month and found an unmarked envelope that contained hundreds of the missing pills, removed from the container they arrived in. Investigators examined the envelope and were able to recover degraded saliva that was used to seal it shut. Comparing the DNA in the saliva to DNA samples from all 50 people who had access to the secure inventory area led investigators to conclude, with [Objective Evidence Strength] % certainty, that the saliva came from Mark as compared to any other employee.
>
> Sample non-legal scenario: Huntington's disease
> A genetic test on a patient reveals that a section of their DNA contains 36 repeats of the 'CAG' sequence. When patients are found to have this many repeats of the 'CAG' sequence it can be concluded, with [Objective Evidence Strength] % certainty, that they will develop Huntington's disease.

The final variable that was manipulated across subjects was the decision criteria instruction, which could take one of three forms. The first included no specific criterion language so as to assess the participants' intrinsic decision criteria in the absence of external guidelines. This condition is referred as the "intuitive belief" (IB) instruction condition. The two other instructions corresponded to the legal burdens of proof of preponderance of the evidence (PoE)

and beyond a reasonable doubt (BaRD). These two burdens of proof were excised from pattern jury instructions adopted by federal courts across the country (U.S. District Court N. D. Cal., 2012) and adapted so that they could be applied to both legal and non-legal contexts. Figure 2 provides sample decision criteria instructions for legal (top row) and non-legal (middle row) contexts.

	Preponderance of Evidence	Beyond a Reasonable Doubt
Legal Context	Start from a presumption that Mark did not steal the drugs. This presumption requires you to conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark stole the drugs by a preponderance of the evidence. That means that the evidence produced leads you to believe that Mark having stolen the drugs is more likely true than not. To put it differently, if you were to put the evidence favoring Mark having stolen the drugs on one side of a balance scale and the evidence favoring Mark not having stolen the drugs on the opposite side, the evidence has to make the scale tip somewhat in order to conclude that Mark stole the prescription drugs.	Start from a presumption that Mark did not steal the drugs. This presumption requires you to conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark stole the drugs beyond a reasonable doubt. Proof beyond a reasonable doubt is proof that leaves you firmly convinced that Mark stole the drugs. There are very few things in this world that we know with absolute certainty, and it is *not necessary* that the proof overcomes every possible doubt. If, based on your consideration of the evidence, you are firmly convinced that Mark stole the drugs, you must conclude that he did. If on the other hand, you think there is a real possibility that Mark did not steal the drugs, you must conclude that he did not.
Non-Legal Context	Start from a presumption that the patient will not develop Huntington's disease. This presumption requires you to conclude that the patient will not develop the disease unless you are satisfied that the facts above proved that the patient will develop the disease by a preponderance of the evidence. That means that the evidence produced leads you to believe that the patient developing Huntington's is more likely true than not. To put it differently, if you were to put the evidence favoring the patient developing the disease on one side of a balance scale and the evidence favoring the patient not developing the disease on the opposite side, the evidence has to make the scale tip somewhat in order to conclude that the patient will develop Huntington's disease.	Start from a presumption that the patient will not develop Huntington's disease. This presumption requires you to conclude that the patient will not develop the disease unless you are satisfied that the facts above proved that the patient will develop the disease beyond a reasonable doubt. Proof beyond a reasonable doubt is proof that leaves you firmly convinced that the patient will develop Huntington's disease. There are very few things in this world that we know with absolute certainty, and it is *not necessary* that the proof overcomes every possible doubt. If, based on your consideration of the evidence, you are firmly convinced that the patient will develop the disease, you must conclude that the patient will develop Huntington's disease. If on the other hand, you think there is a real possibility that the patient will not develop Huntington's, you must conclude that the patient will not develop the disease.
Psychophysical Context	Start from a presumption that the dots are not moving together. This presumption requires you to conclude that the dots are not moving together unless you are satisfied that the dots are moving together by a preponderance of the evidence. This means that you believe that the dots moving together is more likely true than not. To put it differently, if you were to put the evidence favoring the dots moving together on one side of a balance scale and the evidence favoring the dots not moving together on the opposite side, the evidence has to make the scale tip somewhat in order to conclude that the dots are moving together.	Start from a presumption that the dots are not moving together. This presumption requires you to conclude that the dots are not moving together unless you are satisfied that the dots are moving together beyond a reasonable doubt. Proof beyond a reasonable doubt is proof that leaves you firmly convinced that the dots are moving together. There are very few things in this world that we know with absolute certainty, and it is not necessary that the proof overcomes every possible doubt. If, based on your consideration of the evidence, you are firmly convinced that the dots are moving together you must conclude that the dots are moving together. If on the other hand, you think there is a real possibility that the dots are not moving together, you must conclude that the dots are not moving together.

Figure 2. Sample instructions for preponderance of the evidence (PoE) and beyond a reasonable doubt (BaRD). The top and middle rows show the sample instructions from Experiment 1 for the legal and non-legal context respectively. The bottom row shows the instructions for the psychophysical task in Experiments 2 and 3.

Participants selected either "Yes" or "No" in response to a question about their beliefs regarding the scenario. Specifically, those in the intuitive belief condition were asked if they believed that the action or event described in the scenario had or would occur (e.g. "Do you believe that Mark stole the prescription drugs?"; "Do you believe that the patient will develop Huntington's disease?"). For those in either of the burden of proof instruction conditions, the prompt included the specific instruction language within the question (e.g. "Do you believe by a preponderance of the evidence that Mark stole the prescription drugs?"; "Do you believe beyond a reasonable doubt that the patient will develop Huntington's disease?"). This language was included to ensure that participants were incorporating the instructions into their subsequent decisions as they were instructed to do. We avoided using words such as "responsible" or "guilty" to probe participants' beliefs in the legal scenarios in order to keep the prompts comparable across both legal and non-legal contexts, and because pilot data indicated that these words implied certain burdens of proof and consequences which influenced participants' subsequent decisions beyond the scope of the present study.

After providing a yes/no response, participants proceeded to a new screen which asked them to provide their own subjective probability for the event occurring (e.g. "What do you believe is the probability that Mark stole the prescription drugs?"; "What do you believe is the probability that the patient will develop Huntington's disease?"). Participants responded by clicking and dragging a bar along a number line ranging from 0 to 100. We probed the participants' subjective evidence strength as a measure of the probability they were actually considering when making their decision and to determine the extent to which they believed the information provided in the scenario (i.e. the objective evidence strength). Consistent with previous studies comparing subjective and objective probability estimates (e.g. Erev, Wallsten, & Budescu, 1994; Meyniel,

Schlunegger, & Dehaene, 2015), participants tended to overestimate the evidence strength for lower levels of the objective evidence while underestimating the strength for higher levels of the objective evidence (see Supplementary Materials Section 3, Fig. S7). We therefore focus subsequent analyses on the subjective evidence strength, as this was the probability that participants applied when making their decision.

Participants then responded to an attention check question on a new screen to determine whether they had carefully read the scenario and instructions (see Supplementary Section 1). Finally, they provided basic demographic information and were debriefed.

Statistical Analyses

Psychometric functions were used to characterize the likelihood of an affirmative response by evidence strength, context, and instruction type. For the legal context, an affirmative response meant that based on the evidence participants believed that the protagonist Mark had performed the action described in the scenario. For the non-legal context, an affirmative response meant that based on the evidence participants believed that the event in the scenario had occurred or was going to occur (for instance that a patient would go on to develop Huntington's disease). Analyses were completed using the quickpsy package for R version 3.5.3 (Linares & Lopez-Moliner, 2017) together with custom R code.

Psychometric curves for each condition were fit using maximum likelihood methods and the logistic function, which allows the threshold and slope parameters to vary independently of one another (Gilchrist, Jerwood, & Ismaiel, 2005). We assessed the goodness-of-fit for each curve by calculating the deviance, as well as a distribution of deviances from our 1000 bootstrap samples (Linares & Lopez-Moliner, 2017; Wichmann & Hill, 2001a). A p value less than alpha (<0.05;

deviance not within 95% CI from distribution) indicates that the model is not a good fit for the data; all p values were greater than 0.46.

We obtained estimates of the four psychometric parameters from these curves. The threshold parameter is the value of the evidence strength at which an affirmative response becomes more likely (i.e. value of x when y=0.50), while the slope parameter is the slope of the curve at this threshold. Because some conditions did not reach a lower and/or upper asymptote, we instead use the terms lower and upper bounds to describe the predicted value of y at x=0 and x=100 respectively. We generated 95% confidence intervals for the psychometric curve and parameter estimates using 1000 parametric bootstrap samples of our data (Linares, & Lopez-Moliner, 2017, Wichmann & Hill, 2001b). Supplementary Table S4 presents the psychometric parameter estimates and 95% confidence intervals for each condition.

We then compared these parameter estimates between conditions via planned pairwise comparisons of the effect of instruction type within each context level as well as the effect of context within each level of instruction type. For each pairwise comparison, we generated a distribution of difference scores using the 1000 bootstrap estimations from the two parameters (Linares & Lopez-Moliner, 2017, Wichmann & Hill, 2001b), which allowed us to generate confidence intervals with a Bonferroni correction for multiple comparisons (i.e. CI=1-(0.05/9)). Confidence intervals that do not contain zero indicate a significant difference between groups. Supplementary Table S5 presents the differences and Bonferroni-corrected confidence intervals for all pairwise comparisons described below.

Results

Figure 3A shows the psychometric functions of the likelihood of an affirmative response by participants' subjective evidence strength and instruction type for legal and non-legal contexts.

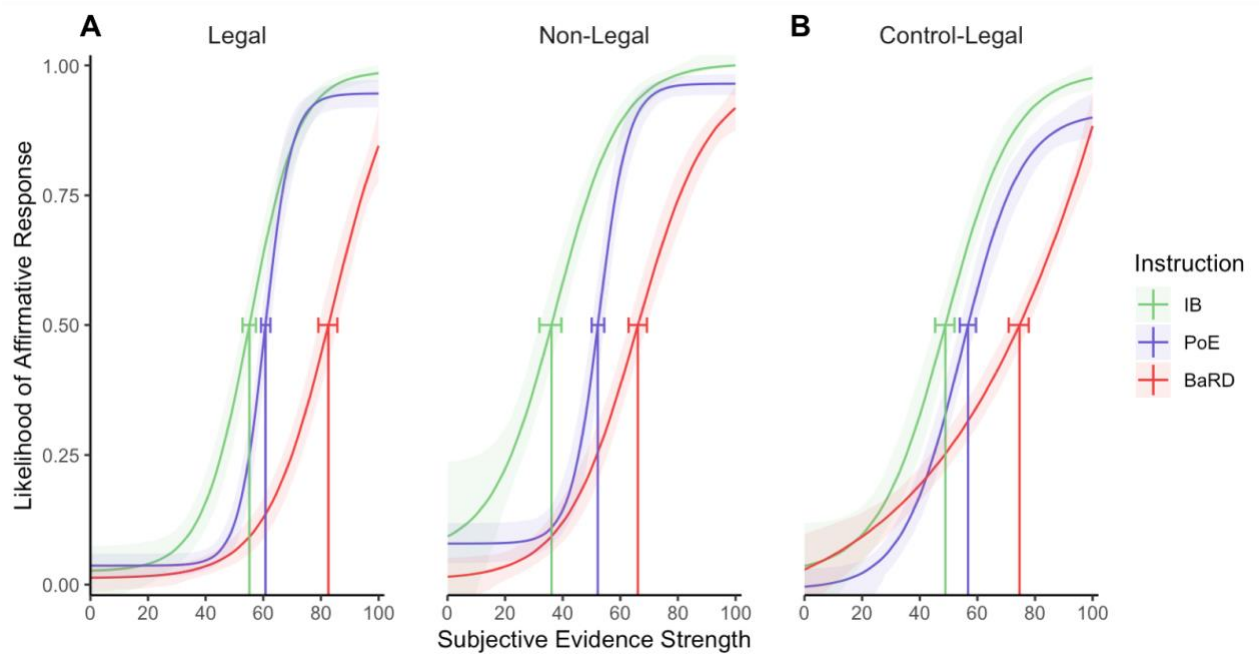

Figure 3. Likelihood of an affirmative response by subjective evidence strength and instruction type for legal and non-legal contexts (A) and the Experiment 1B control-legal scenarios (B). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines.

For both the legal and non-legal contexts there was an effect of decision criteria instructions on participants' decisions such that they were most conservative in the BaRD condition followed by the PoE condition and then the IB condition (see Table S5 for all pairwise comparisons). This was borne out in comparisons of the decision thresholds within both the legal and non-legal contexts, with higher thresholds indicating that participants required stronger evidence to make an affirmative response (Fig. 4A). In both the legal and non-legal contexts, the BaRD threshold was significantly more stringent than both PoE (Legal PoE-BaRD: -21.81[-26.64, -17.11]; Non-Legal PoE-BaRD: -13.91[-19.12, -8.60]) and IB (Legal IB-BaRD: -27.41[-32.34, -21.69]; Non-Legal IB-

BaRD: -30.01[-36.44, -23.26]), and the IB threshold was significantly more lenient than PoE (Legal IB-PoE: -5.60[-9.67, -1.79]; Non-Legal IB-PoE: -16.10[-22.51, -10.87]). The effect of instruction was similarly reflected in the upper bounds of the curves in both contexts (Fig. 4D); IB was greatest (least stringent), followed by PoE and BaRD, with a significant difference between IB and BaRD in both contexts (Legal IB-BaRD: 0.14[0.04, 0.23]; Non-Legal IB-BaRD: 0.08[0.02, 0.18]) and between IB and PoE in the non-legal context (IB-PoE: 0.04[0.01, 0.12]), suggesting that even at higher levels of evidence strength participants are more conservative in their decisions after receiving legal instructions. Together these results suggest that participants' decisions are consistent with the goals of the legal system in that they apply a more conservative decision criterion for BaRD versus PoE, but that they are also more conservative in response to a PoE instruction compared to when they receive no legal instruction (i.e. IB).

Interestingly, the slope estimates followed a different trend in both the legal and non-legal contexts. The PoE slope was steepest followed by IB and BaRD respectively (Fig. 4B). In both contexts, PoE was significantly steeper than both IB (Legal IB-PoE: -0.10[-0.31, -0.01]; Non-Legal IB-PoE: -0.11[-0.28, -0.03] and BaRD (Legal PoE-BaRD: 0.13[0.04, 0.33]; Non-Legal PoE-BaRD: 0.12[0.04, 0.27]. A steeper slope indicates that participants were more responsive to a change in evidence strength, suggesting that for the PoE instruction there was a sharper threshold applied across participants.

We also examined the effect of context (legal versus non-legal) on participants' decisions and found that across instruction type decisions were more lenient in the non-legal versus legal context, as evidenced by the significantly lower decision thresholds in the non-legal condition (Fig. 4A; IB Legal-Non-Legal: 19.03[13.01, 26.03]; PoE Legal-Non-Legal: 8.53[4.69, 12.30]; BaRD Legal-Non-Legal: 16.43[10.33,22.94]). In other words, participants required less evidence to

provide an affirmative response when they assessed non-legal scenarios as compared to legal scenarios.

Finally, we compared the decision thresholds for each condition (instruction x context) to the prescribed legal standards for PoE (50%-55%; McCauliff, 1982; Simon & Mahan, 1971) and BaRD (90%) by determining whether these standards fell within the 95% CI for each threshold estimate (See Fig. 4A and Table S4). Thresholds for the BaRD instruction were significantly lower than the 90% standard in both the legal (82.55, 95% CI[79.09, 85.68]) and non-legal (66.12, 95% CI[62.84, 69.19]) contexts. As for the PoE standard, the PoE threshold was significantly greater than 50% in the legal context (60.74, 95% CI[59.21, 62.40]), while the interval for the non-legal context was just over 50% (52.21, 95% CI[50.05, 54.40]), consistent with the prescribed standard. The IB threshold for the legal context (55.14, 95% CI[52.81, 57.42]) was significantly greater than 50%, while the IB threshold for the non-legal context (36.12, 95% CI[31.95, 39.58] was significantly less than 50%. Thus, both decision standards fell out of the prescribed range in the legal context, whereas those standards were more lenient in the non-legal context.

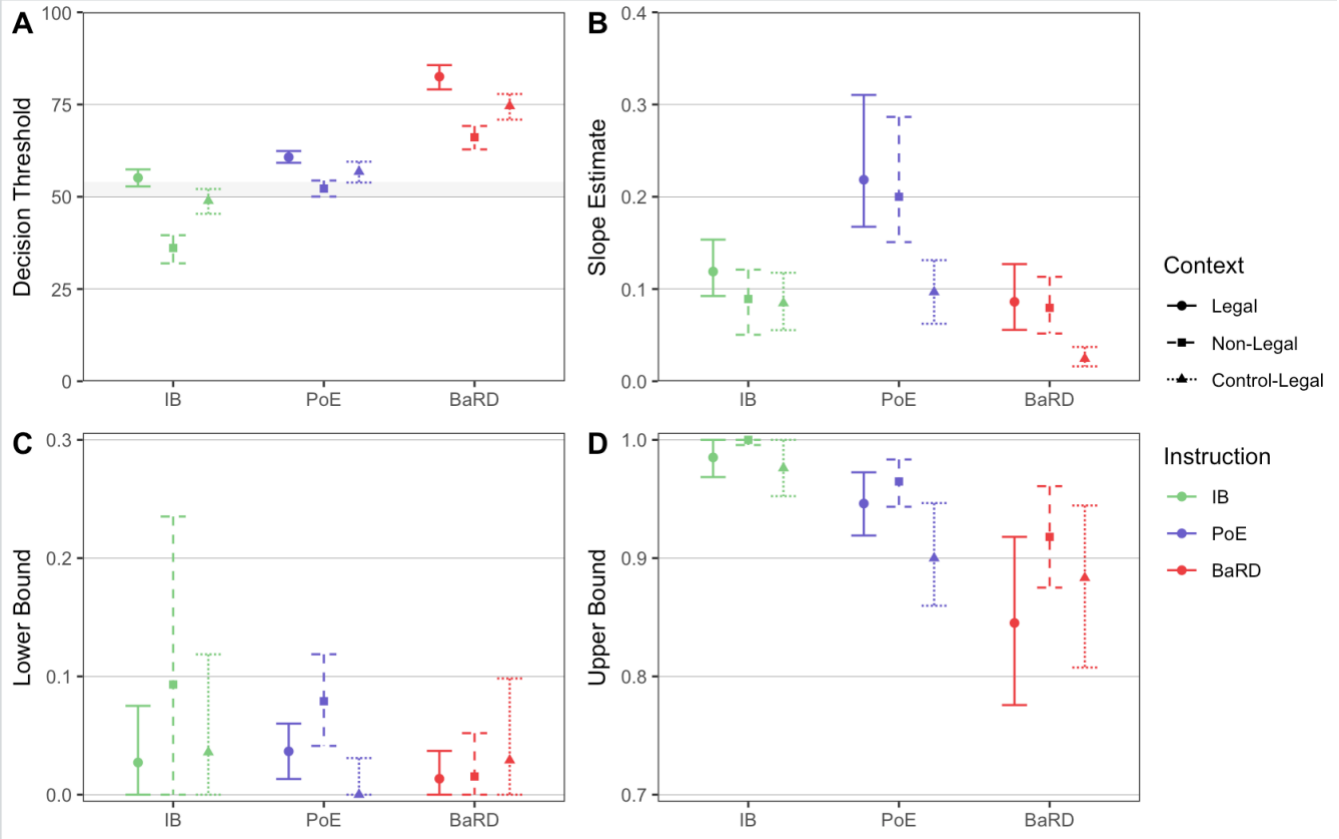

Figure 4. Subjective evidence strength at the decision thresholds (A), slope estimates (B), lower bound (C) and upper bound (D) by instruction type (IB: green; PoE purple; BaRD red) and context for Experiment 1 Legal (solid line w/ circles) and Non-Legal (dashed line w/ squares), as well as Experiment 1B Control-Legal (dotted line w/ triangles). 95% confidence intervals estimated via 1000 bootstrap samples. Gray bar in A shows the 50-55% PoE standard.

Control Experiment 1B

Our finding that decision thresholds were more stringent in the legal context regardless of instruction type may indicate that there is some aspect of the legal context that renders decision-makers more guarded in returning an affirmative decision, perhaps because individuals infer the potentially dire consequences (to the defendant in particular) associated with such decisions. Another possibility, however, is that this context effect resulted from differences in the wording of our legal and non-legal scenarios. One difference was in the length of the scenarios; non-legal scenarios consisted of two sentences while legal scenarios consisted of five to seven in order to establish the 'closed' environment for the wrongdoing. However, in a related study we used the

same legal scenarios but lengthened the non-legal ones to make them comparable in length and found that did not affect the decision outcomes of non-legal scenarios (see Chapter 3 of the present thesis).

Another difference in the scenarios between contexts was the wording of the objective evidence strength. Specifically, in the non-legal scenarios the stated objective evidence strength related to the same event/action that the participants ultimately rendered a decision on (for example, the evidence strength referred to the % probability that a patient's DNA sequence was indicative of a disorder and participants were asked to judge whether the patient has that disorder). By contrast, in the legal scenarios the stated objective evidence strength referred to the level of certainty that the physical evidence associated with the wrongdoing was linked to the protagonist, but the participants' decision focused on whether the protagonist had committed that wrongdoing (for example, the evidence strength referred to the % probability that the saliva was Mark's but the participants were asked to judge whether Mark committed the wrongdoing). While subtle, this textual difference may have led participants in the legal context to believe that Mark did not commit the wrongdoing even though they may have believed that the evidence was linked to him (despite our attempts to phrase scenarios such that the evidence could only have come from the perpetrator). This would result in fewer 'affirmative' responses in the legal context (relative to the non-legal context) at a comparable evidence strength.

To determine whether the greater stringency in the legal context versus non-legal context can be explained by this difference in the scenario language, we ran a control experiment in which we modified our legal scenarios to match the language in the non-legal scenarios in a way that directly linked the evidence strength to the wrongdoing that participants rendered a judgement on.

Methods

We recruited 2993 new participants via Amazon Mechanical Turk. After excluding 452 individuals who failed the attention check, 2541 participants were included in the final analyses which gave us a number of responses for each psychometric function comparable to those in the original experiment with at least 40 responses per cell after all exclusions.

All participants responded to a single modified legal scenario randomly assigned out of 9 possible scenarios, as in the original experiment (3 fact patterns x 3 types of evidence). For these scenarios, the final sentence of the text was modified to present the strength of the objective evidence as the level of certainty that Mark committed the wrongdoing, in order to match the language in the non-legal scenarios (c.f. example below to that of Experiment 1). The experimental design, including the decision criteria instructions, prompts, and subsequent statistical analyses were otherwise identical to those used in the original experiment.

> Sample modified control-legal scenario: Prescription drug theft x DNA evidence
> A hospital has recently found that a large amount of prescription drugs went missing from its secure inventory area. The drugs were documented and videotaped being delivered from the manufacturer to the hospital, and drug inventory staff are searched before and after entering the secure area, leaving investigators puzzled. After hearing of a similar incident at another hospital across the country, investigators checked the trash bins in the secure inventory area over the course of a month and found an unmarked envelope that contained hundreds of the missing pills, removed from the container they arrived in. Investigators examined the envelope and were able to recover degraded saliva that was used to seal it shut. Comparing the DNA in the saliva to DNA samples from all 50 people who had access to the secure inventory area it can be concluded, with [Objective Evidence Strength] % certainty, that Mark stole the prescription drugs.

Results

In comparing the control-legal context to the legal and non-legal contexts within each instruction type (see Fig. S8 and Table S6 for comparison of contexts collapsed across instruction), the control-legal context fell between the legal and non-legal contexts such that decisions were

more lenient than in the original legal context but more stringent than in the non-legal context (Fig. 3B; see Tables S7 and S8 for parameter estimates and pairwise comparisons). This was most clearly borne out in the control-legal decision thresholds (Fig. 4A, dotted line w/ triangle) which were between those of the non-legal and legal contexts for all instruction types. Pairwise comparisons showed that the control-legal threshold was significantly different from both original contexts for the IB instruction (Legal-Legal Control: 6.23[1.23, 11.83]; Non Legal-Legal Control: -12.80[-20.30, -6.02]) and BaRD instruction (Legal-Legal Control: 7.95[1.51, 14.56]; Non Legal-Legal Control: -8.48[-14.60, -2.11]), but not significantly different from either original context for the PoE instruction (Legal-Legal Control: 3.94[-0.35, 8.45]; Non Legal-Legal Control: -4.59[-9.65, 0.57]). Within the control-legal thresholds we observed the same finding as in Experiment 1; IB was significantly more lenient than PoE (IB-PoE: -7.88[-14.43, -2.12]) and BaRD (IB-BaRD: -25.68[-31.67, -18.64]), and PoE was significantly more lenient than BaRD (PoE-BaRD: -17.80[-23.73, -11.43]). Furthermore, the control-legal PoE threshold was significantly greater than the prescribed 50% definition (56.80, 95% CI [53.86, 59.51], just as seen in the original legal context (Fig. 4A).

Together, the findings of this control-legal experiment suggest that presenting the legal scenarios using language that matched that of the non-legal scenarios did result in less stringent legal decisions, but not to the extent that it can fully account for differences between the legal and non-legal contexts. This suggests that the more stringent decisions in the legal context are due, at least in part, to properties specific to the legal domain.

Discussion

Experiments 1 and 1B compared the parameters of psychometric functions to assess the effect of legal burden of proof instructions and context on participants' decisions. This approach

revealed the importance of both decision criteria and context in influencing decision making, as described below.

The type of decision criteria instruction that participants received influenced their decisions in a manner generally consistent with the objectives of the legal system (*In re Winship,* 1970); the PoE burden of proof instruction led to lower, more lenient, decision thresholds than the BaRD instruction, a finding which held true within both the legal and non-legal contexts. The conservativeness of the BaRD instruction was further reflected in the decreased upper bounds for the BaRD instruction conditions (both legal and non-legal), as participants were reluctant to respond affirmatively even at the highest levels of evidence strength (i.e. 100%). While the relative decision thresholds for BaRD and PoE were consistent with the intent of the justice system, the absolute values of these thresholds did not correspond very well to the theoretical values associated with each burden of proof. Specifically, thresholds for the BaRD instruction were significantly lower than the prescribed 90% in both the legal (82.55, 95% CI[79.09, 85.68]) and non-legal (66.12, 95% CI[62.84, 69.19]) contexts. This is consistent with a number of prior studies that have found that individuals tend to interpret the BaRD standard more leniently than the legally prescribed threshold (Dhami et al., 2015; Horowitz & Kirkpatrick, 1996; Simon & Mahan, 1971; Teitcher & Scurich, 2017). By contrast, the PoE instruction was overly stringent (i.e. significantly greater than 50%) within the legal context (60.74, 95% CI[59.21, 62.40]), while not significantly different from 50% in the non-legal context (52.21, 95% CI[50.05, 54.40]). Interestingly, Simon and Mahon (1971) also found that jurors' estimates for the PoE standard were more stringent than 50%, instead averaging around 75%. In that study, participants were asked to quantify PoE in regards to a decision to convict, which implied a criminal legal context. It is therefore possible that while our legal scenarios described acts that could be considered either of criminal or civil

wrongdoings (data theft, drug theft, murder), participants viewed them as more criminal than civil. Thus the overly stringent PoE standards observed in this and the previous study (Simon & Mahon, 1971) may be driven in part by participants inferring a criminal context. The non-legal PoE threshold was consistent with the prescribed 50% standard, possibly because there was no implication of any such criminal context. However, it is not clear whether decisions in response to strictly civil legal scenarios would align more closely with the legal or non-legal scenarios. A future study can address this issue by conducting a control experiment in which participants respond to the PoE instruction for a legal scenario in a civil context (i.e. protagonist faces litigation for wrong-doing); if this PoE threshold is significantly more lenient than the legal PoE threshold in the present study, it would provide support for the idea that the overly stringent application of PoE in the legal context is due in part to participants inferring a criminal context.

How the PoE and BaRD legal burdens of proof relate to decision-making in the absence of any legal instruction had not been explored previously. We found that the threshold for IB decisions was lower than both the PoE and the BaRD instruction decisions regardless of context. This is somewhat surprising given that PoE theoretically reflects a very lenient decision standard of about 50%. It is also inconsistent with the hypothesis that the overly stringent interpretation of PoE and overly lenient interpretation of BaRD in previous studies (e.g. Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; McCauliff, 1982; Simon & Mahan, 1971) is due to individuals shifting each of the standards to align more closely with an intuitive decision threshold that falls somewhere between them. These results suggest that individuals' intuitive decision standards may simply be quite liberal, to the point – at least in the non-legal context –of providing an affirmative response even when the evidence was below 50%. Conversely, these findings imply that as soon as decision-making involves a legal context, either because the scenario

describes a legal situation *or* the decision criterion is a legal standard, participants adopt a more stringent decision criterion.

A common finding across instruction types is that legal scenario cases were adjudicated more conservatively than non-legal scenarios. Experiment 1B suggests that this difference cannot be accounted for just by differences in the wording of legal and non-legal scenarios. Rather, we surmise that this difference is due to participants inferring negative consequences for the perpetrator in the legal context, despite our attempt to keep the prompt language neutral (i.e. No use of "guilty" or "responsible"). Given that a crime took place in the legal context and was paired with legal burden of proof instructions for the PoE and BaRD conditions, it is not unreasonable for participants to assume that the scenario's protagonist could face charges and punishment based on their decision. Indeed, previous studies have found that the type of charge against an individual as well as punishment outcomes can influence conviction decisions (Kerr, 1978; Vidmar, 1972). Conversely, scenarios in the non-legal context may have favored neutral or affirmative responses. For example, false positives could be preferable to false negatives when dealing with predicting Huntington's disease in order to facilitate treatment, leading to decreased thresholds across instruction type for non-legal contexts. This could further account for the non-legal IB threshold falling below 50%.

The upshots of Experiments 1 and 1B are that 1) individuals are increasingly more stringent for the PoE and BaRD standards than when they receive no instruction, a finding that is inconsistent with the prevalent notion that PoE and BaRD represent extreme boundaries of decision criteria on opposite sides of people's general intuitions, and 2) that legal scenarios are adjudicated more stringently than a range of non-legal scenarios. But to what extent do these results generalize beyond our experimental scenarios? Both legal and non-legal domains relied on

scenarios which presented the objective evidence strength as a frequentist probability, and the differences between domains suggest that scenario content influences how individuals apply legal decision standards. In our subsequent experiments we pushed the boundaries of the applicability of our results by assessing their generality in a completely distinct domain; psychophysics.

Experiment 2

In Experiment 2, we sought to assess the extent to which the results from Experiment 1 are generalizable by assessing the effect of legal decision criteria instructions in a classic perceptual task, namely a dot motion coherence task. This task has been used extensively to study decision processes using psychometric functions (e.g. McGovern, Roach, & Webb, 2014; Pilly & Seitz, 2009; Van Wezel & Britten, 2002). Were we to find comparable effects of legal instructions from Experiment 1 in this psychophysical domain, this would suggest that our findings generalize across a broad range of decision-making domains, from the perceptual world to high-level cognition. This would also lend further validity to the use of psychometric functions to assess complex decision-making. In Experiment 2, participants completed a single block of a dot motion task with assignment to one instruction type (IB, PoE, BaRD).

Methods

Participants

We recruited 72 participants to complete the experiment (26 males, mean age=20.45 years) in the laboratory. All participants provided informed consent, and the experimental protocol was approved by the Vanderbilt University Institutional Review Board. Participants provided basic demographic information and were debriefed at the end of the study. They received either course

credit for an introductory psychology course or were paid $12 per hour for completing the task, which took approximately 30 minutes. Twenty-four of the participants emanated from a pilot study in which participants completed 3 blocks, one of each instruction type, but always beginning with IB followed by PoE and BaRD in random order. As we observed an order effect for the presentation of the two legal standards in the pilot study, we chose to present only a single instruction block for the other 48 participants. We included the IB block data from the 24 participants who completed the pilot study because they all completed the IB block first and were not aware of the PoE and BaRD instructions until after they had completed the first (IB) block. The results presented below are qualitatively similar when excluding these 24 participants. One participant was excluded due to their needing to leave before completing the entire task, leaving 71 for the analyses described below.

Materials and Design

Participants completed a random dot motion task in which they were asked to report whether or not they perceived coherent motion within a set of moving dots. The task employed a 3 (decision criteria instruction, between-subjects: IB, PoE, BaRD) x 6 (evidence strength, within-subjects; 0%, 20%, 40%, 60%, 80%, 100%) design. The evidence strength corresponded to the percent coherence of the dots, or the proportion of dots moving together in the same direction. Participants completed 30 trials at each coherence level in random order for a total of 180 trials.

Dot Task Parameters. The task was presented on a computer screen using custom code via Psychtoolbox Version 3 for MATLAB R2017. Each trial consisted of a random dot motion display. Fifty white dots, each 8 x 8 pixel squares, were presented on a black background within an invisible circular aperture with a 12-degree diameter centered on the screen. Dots were randomly distributed within the circular aperture at the beginning of each trial. The evidence

strength/coherence level was randomly assigned for each trial; the proportion of dots moving in the same direction were assigned to move either up or down while the remaining dots were assigned a random direction of movement. Dots moved 5 degrees per second in their assigned direction and had a limited lifetime of 10 frames with a frame rate of 60 Hz. Each dot was assigned a random starting frame value between 0 and 9, and jumped to a new random position within the aperture once this value reached 10, which prevented subjects from tracking the motion of individual dots. Dots that moved outside of the aperture were similarly assigned to a new random position within the aperture to ensure that dot density remained constant.

Each random dot motion trial was displayed for 500 milliseconds followed by a 10 second response window with the question prompt (described below). Once participants responded, there was a 1 second inter-trial interval during which a 0.5° white fixation square appeared in the center of the screen, after which the next trial commenced.

Pilot data indicated that these task parameters allowed subjects to reliably identify the 0% and 100% coherence trials while struggling with the intermediary strength levels. This produced psychometric curves that could be compared across instruction type.

Task Instructions. Participants first received verbal instructions describing the task. They were told that they would view a set of moving dots on each trial and would be asked if they believed the dots were moving together. "Moving together" was defined as being able to perceive coherent motion within the dot aperture, either up or down. Participants were told that on each trial they might see all, some, or none of the dots moving together, and that not all of the dots had to move in the same direction for them to perceive coherent motion. These verbal instructions were then reiterated in writing on the computer screen at the start of the task for participants to read at their own pace.

Practice Trials. Prior to completing the experimental block, participants completed a set of practice trials in order to adapt to the task. Specifically, they first viewed a trial with 100% coherence and were explicitly told that this was a sample of all of the dots moving together, followed by a trial with 0% coherence and a statement that this sample represented a case where none of the dots were moving together in the same direction except by chance. Participants then completed 10 additional practice trials with no feedback for a total of 12 practice trials, such that they viewed each coherence level twice (presented in random order) prior to the experimental block.

Decision Criteria Instructions. Participants were randomly assigned to one of the three decision criteria instruction levels (IB, PoE, BaRD). Those in the IB condition received no additional instructions after the initial description of the task and practice trials. Participants in the PoE and BaRD conditions read the burden of proof instructions on the screen immediately after completing the practice trials. Figure 2 (bottom row) includes the specific language used for both instructions for the dot motion task.

Subject Responses. After each trial, participants were asked to provide a yes/no response as to whether or not they believed that the dots were moving together (i.e. "Do you believe that the dots were moving together?"; "Do you believe by a preponderance of the evidence that the dots were moving together?"; "Do you believe beyond a reasonable doubt that the dots were moving together?"). Participants responded using the keyboard, with "f" for yes and "j" for no.

Statistical Analyses

The methods used to generate the psychometric curves and the subsequent statistical analyses were identical to those used in Experiment 1. We assessed the goodness-of-fit using the deviance and distribution of bootstrap deviances. All p values were greater than 0.23 indicating

that the curves were a good fit for the data. Pairwise comparisons were again performed by generating a distribution of difference scores using the bootstrap estimations from the two parameters (Linares & Lopez-Moliner, 2017, Wichmann & Hill, 2001b) to generate confidence intervals with a Bonferroni correction for multiple comparisons (i.e. CI=1-(0.05/3)). Confidence intervals that do not contain zero indicate a significant difference between groups.

Results

Figure 5 presents the psychometric curves for Experiment 2. Supplementary Table S9 presents the parameter estimates with 95% confidence intervals for each condition. Supplementary Table S10 presents the differences and Bonferroni-corrected confidence intervals for all pairwise comparisons described below.

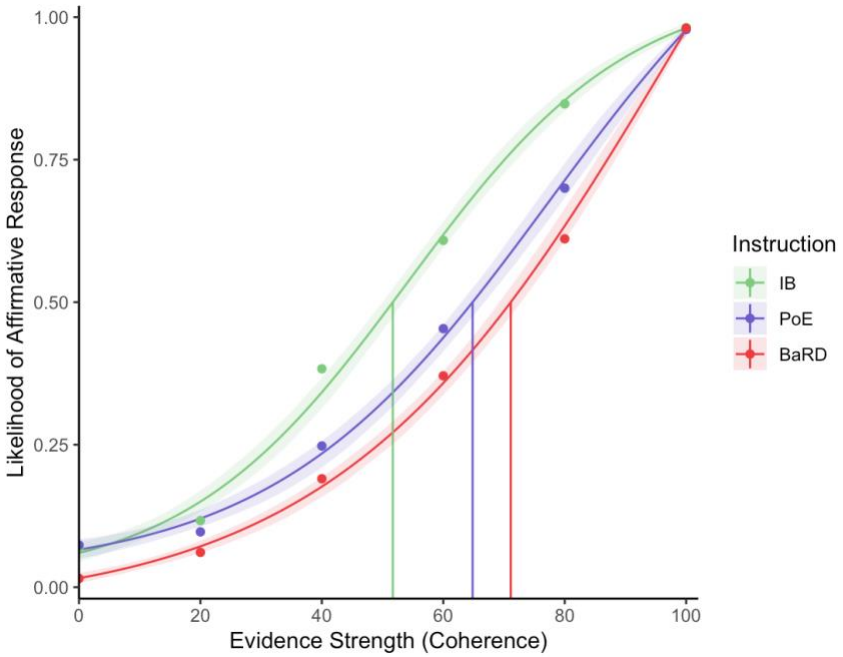

Figure 5. Likelihood of an affirmative response by evidence strength (% coherence of dots) and instruction type for Experiment 2. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines.

Consistent with Experiment 1, there was an effect of decision criteria instructions on participants' decisions such that they were most conservative in the BaRD condition, followed by PoE and then IB (Fig. 6A). Comparisons of the thresholds were significant between all instruction levels (IB vs BaRD: -19.43[-22.27, -16.47]; IB vs PoE: -13.14[-16.16, -10.21]; PoE vs BaRD: -6.29[-9.43, -3.00]. Participants required stronger evidence (greater % coherence) to state that the dots were moving together when they received legal standard instructions, and this was more pronounced for the BaRD instruction versus the PoE instruction. The lower bounds also demonstrate the more liberal decisions in the IB and PoE conditions compared to BaRD (Fig. 6C), as the BaRD estimate was significantly less than those for IB (IB vs BaRD: 0.04[0.02,0.07]) and PoE (PoE vs BaRD: 0.05[0.03,0.07]).

As in Experiment 1, we compared the decision thresholds for each condition to the prescribed legal standards for PoE (just above 50%) and BaRD (90%) by determining whether the standards fell within the 95% CI for each condition (see Table S9). The BaRD threshold was significantly less than 90% (71.11, 95% CI[69.24, 72.81], and the PoE threshold was significantly greater than 50% (64.83, 95% CI[62.99, 66.78]). The IB threshold was consistent with the PoE standard (i.e. not significantly different from just above 50%; 51.69 95% CI[50.01, 53.43].

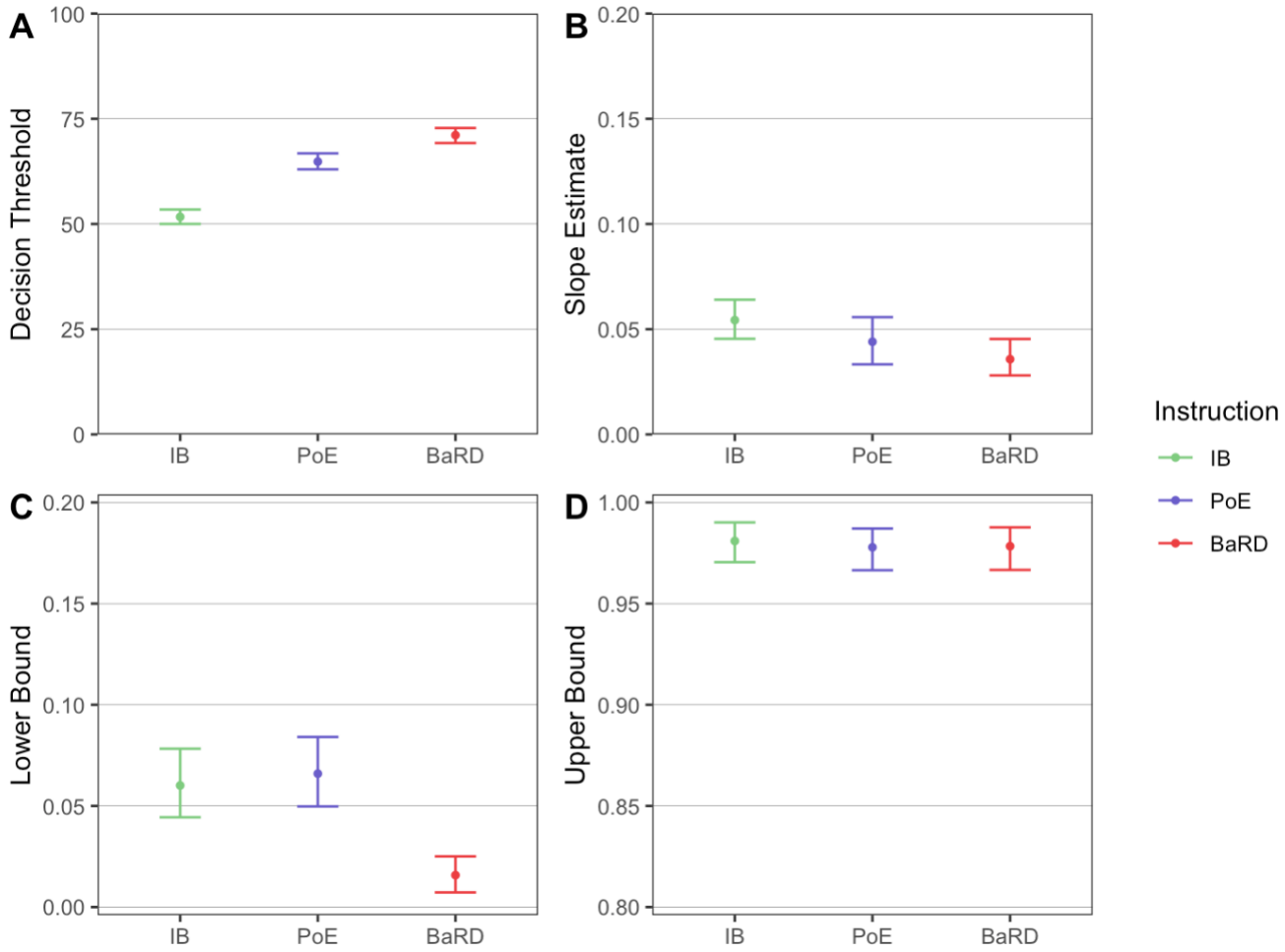

Figure 6. Evidence strength (% coherence) at decision thresholds (A), slope estimates (B), lower bounds (C), and upper bounds (D) by instruction type for Experiment 2. 95% confidence intervals estimated via 1000 bootstrap samples.

In a final analysis, we compared decision thresholds in the dot motion coherence task to those of the legal context in Experiment 1 (see Table S11 for all parameter comparisons). We found that the BaRD decision threshold was more conservative in the legal domain than in the perceptual domain (Exp1 Legal-Exp2: 11.44[7.00, 15.60]), while the PoE threshold was more conservative in the perceptual domain (Exp1 Legal-Exp2: -4.09[-7.17, -1.24]). The IB thresholds were not significantly different (Exp1 Legal-Exp2: 3.46[-0.27, 6.94]). By contrast, comparison of the dot motion coherence task with the non-legal condition of Experiment 1 showed that decision thresholds were significantly more conservative for the perceptual domain than the non-legal

across all instructions (see Table S12 for all parameter comparisons; IB: -15.57[-21.27, -11.08]; PoE: -12.62[-16.19, -9.02]; BaRD: [-9.50, -0.33]).

Discussion

Using a classic psychophysical motion coherence task in Experiment 2, we were able to replicate the two main findings derived from the legal and non-legal scenarios of Experiment 1: There was a systematic order effect of legal instructions on decision thresholds, and participants generally adopted a more stringent decision criterion in the legal than in the perceptual domain.

First, with respect to the effect of legal instructions, we observed in Experiment 2 that the IB condition had the most lenient decision threshold followed by PoE and BaRD, consistent with our findings in Experiment 1, suggesting that the application of these legal standards is generalizable. Here again, IB is less stringent than PoE even after removing the potential influence of the scenarios. These results also suggest that participants are overly stringent in response to the PoE instruction as the PoE threshold was well above 50%. This may mean that the PoE language itself leads to more stringent decision thresholds, perhaps because it connotes a legal context (but see below). Additionally, McCauliff (1982) found that a sample of judges rated "preponderance of the evidence" as near 50-55%, but as higher than "more probable than not", which would also be consistent with the idea that the phrase "preponderance of the evidence" increases decision thresholds despite being theoretically equivalent to the idea of more probable than not. It is also interesting to note that it is once again the IB condition – not PoE – that hovers around a 50% decision threshold.

Second, decisions were adjudicated less stringently in the perceptual than in the legal domain, at least under BaRD instructions. This replicates the finding in Experiment 1 that legal

decision-making is held to a higher standard than non-legal decision-making, although conversely the PoE instruction led to more conservative decisions in the perceptual versus legal domain. A potential explanation for this oddity is that participants interpreted the PoE instructions to mean that a preponderance of the dots had to move together and that they perceived this to only to be the case when about 65% of the dots did so. Differences in the spread of the evidence strength between experiments may also contribute to this difference. More experiments will be necessary to account for this finding.

Aside from the similarities, Experiments 1 and 2 also revealed notable differences. In particular, Experiment 2 did not exhibit the same effect of instruction on the upper bounds that was present in Experiment 1. This may speak to participants' degree of certainty in the evidence provided. The dot stimuli parameters were selected because a pilot study found that they allowed participants to consistently identify the 0% and 100% coherence levels, in order to generate psychometric curves. Individuals also completed multiple trials which gave them a sense of the different coherent levels and allowed them to recognize the strongest evidence level (i.e. 100%) with confidence. Participants did show evidence of guessing on the 0% coherence trials in the IB and PoE condition, possibly because they perceived some amount of coherence within the purely random motion and were more lenient in these conditions.

While the results of Experiment 2 were broadly consistent with those found using more complex decision processes in Experiment 1, we considered whether these similarities were fortuitous outcomes of the methodological differences across experiments. Specifically, the dot motion task presented multiple trials to each subject for any given instruction type, whereas Experiment 1 presented only a single trial to each subject. Therefore, in Experiment 3 we presented a single dot motion trial to each subject to more closely align with the design used in Experiment

1. If methodological differences are not critical, the results of Experiment 2 should be replicated in Experiment 3.

Experiment 3

In Experiment 3 we combined the single-trial approach of Experiment 1 with the random dot motion task of Experiment 2. Participants were therefore exposed to a single decision criteria instruction type and responded to a single dot-motion task trial.

Methods

Participants

We recruited 1212 participants from the United States via Amazon Mechanical Turk (56% male, mean age=36.82 years). We excluded 18 participants who indicated that they experienced technical issues and were unable to view all of the clips. We further excluded those who did not correctly answer the two attention check trials, leaving a total of 786 participants in subsequent analyses. As in Experiment 1, this allowed us to have forty participants per condition after all exclusions were applied. Participants received $0.40 for completing the task, which took less than five minutes on average. All participants provided informed consent, and the experimental protocol was approved by the Vanderbilt University Institutional Review Board. Participants provided basic demographic information and were debriefed at the end of the study.

Design and Materials

Participants completed a single experimental trial of the random dot motion task described in Experiment 2. The task employed a 3 (decision criteria instruction; IB, PoE, BaRD) x 6 (evidence strength; 0%, 20%, 40%, 60%, 80%, 100%) between-subjects design. As in Experiment

2, the evidence strength corresponded to the percent coherence of the dots, or the proportion of dots moving together in the same direction.

The experiment was administered using the Qualtrics online survey platform. Participants were required to use a computer to complete the survey- the survey screened out individuals who were using a mobile device. This was done to ensure that participants could successfully view the dot stimuli, which were presented as video clips. Participants began by reading the same general task instructions presented in Experiment 2 that described the goal of identifying coherent motion within the dots. Participants completed practice trials before completing a single experimental trial with random assignments to the decision criteria instruction, the coherence level (evidence strength), and the direction of the coherent motion.

Practice Trials. Prior to the experimental trial, participants completed 12 practice trials to acclimate them to the task and allow them to see the full range of motion possible. As in Experiment 2, the first trial was an example of 100% coherence with explicit instruction that it was a sample of all of the dots moving together and the second trial explicitly presented a sample with 0% coherence. The responses to these two trials additionally served as attention checks as participants were told beforehand whether none or all of the dots were moving together. The remaining 10 trials were presented in random order such that participants viewed two samples of each level of the evidence strength over the course of the practice trials. Participants provided a yes/no response for each practice trial on a new page to indicate if they believed there was coherent motion.

Decision Criteria Instructions. Following the practice trials, participants in the PoE and BaRD instruction conditions received the additional legal standard instructions to apply to their

decision which were identical to those used in Experiment 2 (see bottom row of Fig. 2). Participants in the IB condition received no additional instruction following the practice trials.

Dot Task Parameters. The dot task parameters were identical to those used in Experiment 2. The trials were presented as video clips that were 500 milliseconds each in length. Participants were instructed to view each of the video clips only once.

Subject Responses. participants provided a yes/no response as to whether or not they believed that the dots were moving together at the end of each practice trial and at the end of the single experimental trial. Participants who responded "yes" to perceiving coherent motion were also asked to report whether they perceived the dots moving up or down (though the latter report only served to show that participants were more likely to report the direction incorrectly for lower levels of coherence, suggesting that they were guessing under these conditions).

After completing the experimental trial, participants were asked whether they had experienced any technical issues in viewing the video clips for the trials. They then provided demographic information and were debriefed.

Statistical Analyses

The statistical analyses were identical to those in Experiment 2.

Results

Figure 7 presents the psychometric functions for the online dot task by instruction type. We calculated the deviance for each curve to determine goodness of fit. All p values were greater than 0.99, indicating that the model was a good fit for the data. Supplementary Table S13 presents the psychometric parameters and 95% confidence intervals for each condition. Supplementary

Table S14 presents the differences and Bonferroni-corrected confidence intervals for all pairwise comparisons described below.

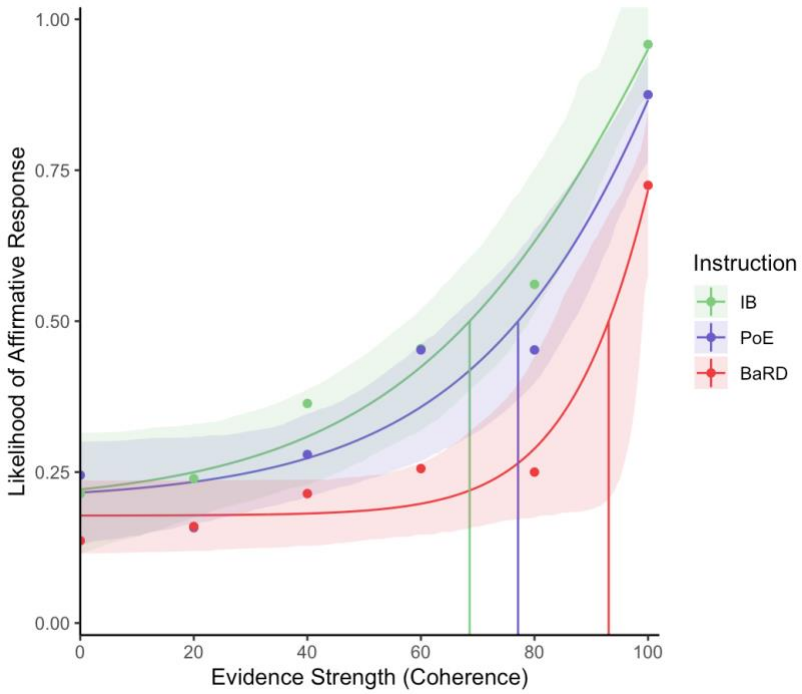

Figure 7. Likelihood of an affirmative response by evidence strength (% coherence) and instruction type for Experiment 3. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines.

The effect of decision criteria instruction showed the same pattern observed in Experiments 1 and 2. BaRD was most stringent, followed by PoE and then IB (Fig. 8A), with a significant difference between the decision thresholds for the IB and BaRD instructions (IB-BaRD: -24.49[-40.00, -6.64]) as well as the upper bound for the IB and BaRD instructions (Fig. 8D; IB-BaRD: 0.23[0.07, 1.04]. In general, the logarithmic shape of the fits suggests that participants were conservative in their response until the highest coherence levels, particularly for the BaRD instructions.

Finally, we compared decision thresholds in the dot motion coherence task to those of the legal condition in Experiment 1 (see Table S15 for all comparisons). Unlike in Experiments 1 and

2, here we found that the decision thresholds were more lenient in the legal domain than in the perceptual domain across all three instruction types, though this was significant only for the PoE instruction (Exp3-Exp1 Legal: 16.37[0.95, 28.56]. Similarly, the same relationship held for the comparison between the present decision thresholds and those in the non-legal domain of Experiment 1 (see Table S16 for all comparisons; IB: 32.46[18.42, 46.50]; PoE: 24.90[8.99, 37.94]; BaRD: 26.94[15.20, 34.77]), or relative to Experiment 2 for that matter (see Table S17 for all comparisons; IB: 16.89[2.87, 29.76]; BaRD: 21.95[10.65, 28.82]).

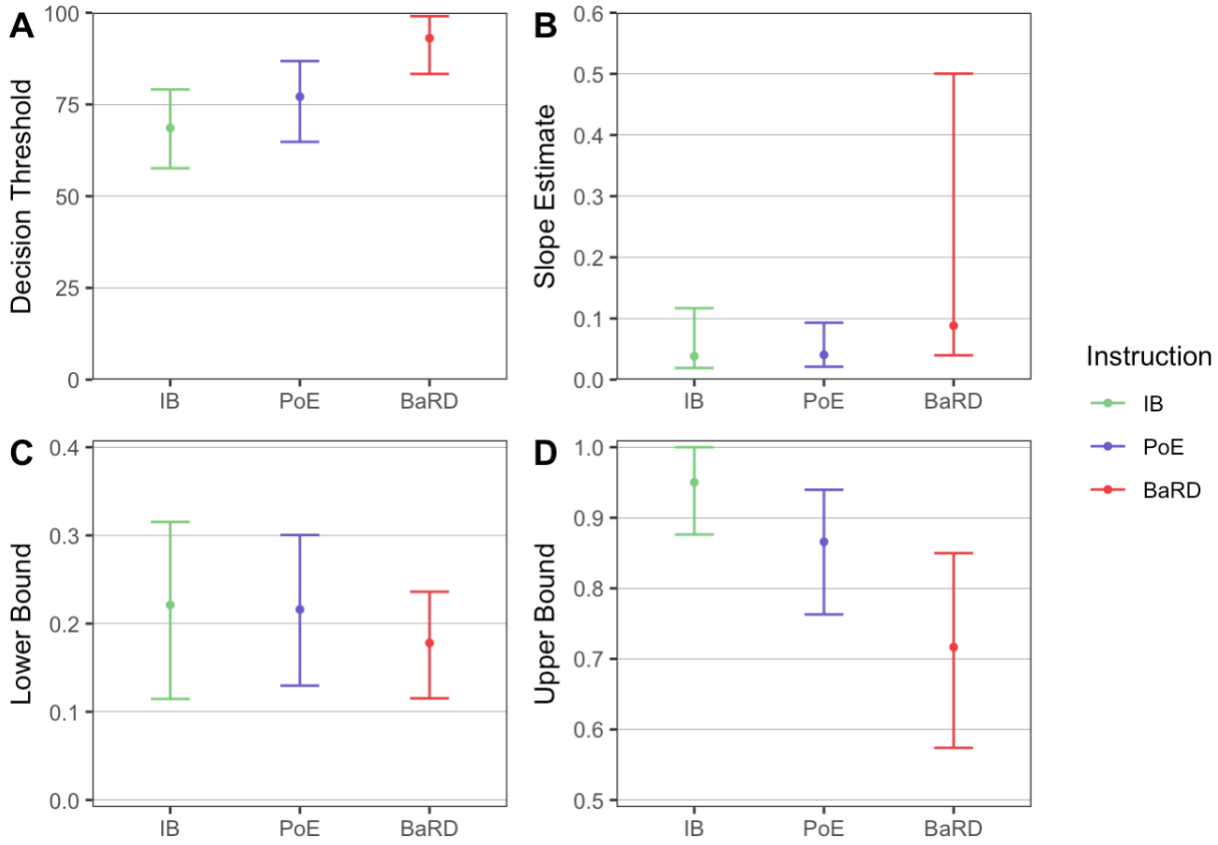

Figure 8. Evidence strength (% coherence) at decision thresholds (A), slope estimates (B), lower bounds (C), and upper bounds (D) by instruction type for Experiment 3. 95% confidence intervals estimated via 1000 bootstrap samples.

Discussion

The effect of decision instructions observed in the previous experiments held true in Experiment 3, with BaRD as more stringent than PoE and IB, and PoE being more stringent than IB. This effect was observed despite the higher data variance associated with this experimental design, which was expected given that each subject's data is based on a single trial when standard perceptual experiments typically contain tens if not hundreds of trials. This variance is also reflected in the relatively high proportion of affirmative responses when the coherence level was 0% (lower bound), showing that many participants were guessing at the lower coherence levels. While this limited our ability to identify differences in some of the parameters (see below), this study nevertheless supports the findings of Experiment 1 by providing similar observations about the order effect of instructions on decision thresholds, even in single trials.

While Experiment 3 generally replicated the findings of Experiment 1 and 2 with respect to the effect of legal instructions, that was not the case in regards to the relative stringency of decision criteria in the legal and perceptual domain. Specifically, the decision thresholds were more conservative in Experiment 3 compared to not only the legal scenario experiment (Experiment 1) but also the previous perceptual experiment (Experiment 2). The exponential shape of the fits suggests that participants were conservative in their response until the highest levels of coherence. Even at those levels, participants were still conservative relative to Experiment 2, in which participants completed the same dot coherence task but had multiple trials. This was shown in Experiment 3's greater thresholds and decreased upper bound estimates compared to Experiment 2 (Table S17 presents the pairwise comparisons between Experiment 3 and 2). A similar pattern was seen when comparing Experiment 3 to the scenarios in Experiment 1 (see Tables S15-S16). The greater lower bounds (suggesting a higher guess rate) together with increased conservativeness at the highest coherence levels suggest that the task was difficult for

participants and that they were less certain in their decisions. This is corroborated by anecdotal evidence from participants' comments that indicated they found the task to be challenging due to having only a single experimental trial to respond to. Evidently, participants adopted a conservative decisional stand as a result of their inexperience with this perceptual task owing to their exposure to a single experimental trial, even with two practice trials per level.

Chapter 1 Discussion

We implemented a psychometric approach to assess the effect of legal standards of proof (i.e. PoE, BaRD) to determine whether they have the intended effects on decision-making, and to determine how such burdens of proof compare to people's intuitive decisions. We further evaluated the generality of these instructions by comparing decisions across three contexts- legal, non-legal, and psychophysical. There are three main upshots of these experiments: 1) a psychometric analysis proved to be a powerful approach to investigate complex decision standards within and across contexts; 2) our intuitive decision standards are more liberal than not only the BaRD standard but also PoE; and 3) legal decision standards are more stringent than non-legal (e.g. medical, scientific and psychophysical) judgments.

To our knowledge, this research is the first to use a psychometric approach to assess higher-level decisions. Critically, a psychometric approach provided a common analytical approach so that we were not only able to compare and quantify the application of legal standard instructions between scenario contexts (i.e. legal versus non-legal), but also were able to compare these decisions with those made in a simple perceptual psychophysics task. Observing the same effect of instruction across such different domains (i.e. IB < PoE < BaRD) indicates that this finding is

robust and not due simply to methodological considerations such as differences in scenario content or providing a frequentist probability estimate (as in Experiment 1).

Furthermore, properties of the different psychometric parameters illuminate multiple aspects of decision-making beyond what can be assessed looking at single measures of decisions. Our decision thresholds provide an estimate of the evidence strength needed to make a "yes" response more likely than a "no" response (i.e. evidence strength where y=50%). These decision thresholds were useful in demonstrating the relative leniency/stringency between our different conditions (particularly for more ambiguous strength levels), and also provide an estimate of accuracy for the application of the legal standards as we assessed whether the decision thresholds were consistent with the prescribed PoE (just over 50%) and BaRD (90%) standards. The slope estimates on the other hand provide an estimate of the strength of the relationship between the independent and dependent variable, with a steeper slope indicating that less of an increase in evidence strength is needed to result in a decision change. The PoE instruction in the legal context of Experiment 1 provides a clear example of how these parameters inform different aspects of decision-making processes; while the decision threshold was significantly different from 50% (meaning participants applied that decision standard more stringently than its intended application), the slope was steeper than both the IB and BaRD instructions (meaning subjects' decisions were more responsive to changes in evidence under PoE than the other standards). The latter finding is in line with the concept that the PoE standard refers to a tipping point between two decisional outcomes even though we found that tipping point to be more stringent than prescribed by the judicial system (McCaulliff, 1982; Simon & Mahan, 1971). The lower and upper bounds of the psychometric function, on the other hand, can reveal limitations in the relationship between evidence and decisions, perhaps best exemplified by the diminutive upper bound of the BARD

instruction in legal scenarios of Experiment 1, thus revealing how BaRD instruction can lead to conservative decisions even under absolute (100% certainty) conditions (see below for more discussion). Evidently, the application of the psychometric approach to legal decision-making provides a powerful tool to dissect complex decision processes.

A major thrust of this study consisted in assessing how decision parameters for the BaRD and PoE instructions compared not only to each other but also to participants' decisions in the absence of any legal instruction (IB). Across both legal and non-legal contexts as well as the psychophysical dot motion tasks, decision thresholds (at 50%) were greatest for the BaRD instruction, followed by the PoE instruction and lastly by the IB condition. That the BaRD thresholds were more stringent than the PoE thresholds is consistent with the ideal distinction put forth by the legal system (Laudan, 2003) as well as with previous work quantifying these burdens (ex. Simon & Mahan, 1971). This appears to be a robust trend as it held true across methodologies and contexts in the present study.

Another finding that was consistent with prior work (e.g. Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; McCauliff, 1982; Simon & Mahan, 1971) is that our decision threshold estimates were more stringent than the (legally) prescribed ones for PoE and more liberal than the prescribed ones for BaRD. PoE is meant to fall around a "tipping of the scales" at just above 50%. However, previous studies have found that jurors and judges alike tend to interpret PoE as having a threshold greater than 50%, with estimates for judges ranging between 50-55% and jurors estimating PoE as high as 75% (McCauliff, 1982; Simon & Mahan, 1971), and our findings from Experiment 2 and 3 suggest that even in the absence of the potential influences of legal scenarios (e.g. participants assuming certain legal outcomes based on their decision) individuals are overly stringent in applying the PoE threshold.

By contrast, we found that the decision thresholds for the BaRD instruction were more lenient than the prescribed 90% threshold, consistent with previous studies (McCauliff, 1982; Simon & Mahan, 1971). Our experiments did not have particularly high stakes however; it may be the case that actual jurors would apply a more stringent BaRD threshold when dealing with a real legal case with actual consequences. However, the PoE interpretations were overly stringent despite the lack of real-life consequences, so it seems unlikely that this alone accounts for the discrepancy. One notable exception to the overly lenient application of the BaRD standard was participants' conservative response to this instruction when the evidence was seemingly irrefutable (i.e. 100%), regardless of scenarios (legal or non-legal) or domains (cognitive or perceptual) tested. This finding suggests that people may have an inherent aversion to accepting absolute levels of proof. In that respect, the (few) percentage points above the BaRD upper bound may be an estimate of the 'unreasonable' doubt.

The overly stringent interpretation of PoE and overly lenient interpretation of BaRD observed in this study and in previous ones (e.g. Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; McCauliff, 1982; Simon & Mahan, 1971) has been interpreted as being a result of individuals shifting each of the standards to align more closely with an intuitive decision threshold that falls somewhere between them. Strikingly, however, we found that participants' IB thresholds were more lenient than the PoE decision thresholds. This is surprising given that Participants' IB decision thresholds fell closer to the 50% theoretical standard than PoE decision thresholds for the legal context in Experiment 1 as well as in Experiments 2 and 3, which suggests that intuitive decision thresholds may naturally reflect the idea of "more likely true than not" more faithfully than the intended PoE instructions for many binary decisions.

Another robust finding of the present study is the systematic effect of context on decisions-making, as decision thresholds were consistently greater for the legal compared to the non-legal context across instruction type in Experiment 1. We ruled out that his effect may be due to subtle differences in scenario phrasing (Experiment. 1B). Rather, we hypothesize that the distinctive stringency of legal decisions is due in part to participants implicitly attaching a punishment to a culpability verdict, which may increase the decision threshold due to the inferred negative consequences that the decision would have on the defendant. Previous studies have found that conviction rates are lower when the defendant faces more severe charges and potential punishment (Kerr, 1978; Vidmar, 1972). Participants may be more averse to a false positive than a false negative in the legal context; indeed, the BaRD standard is founded on this very principle, favoring false acquittals to false convictions at a 10:1 rate (Laudan, 2003; Newman, 1993). Furthermore, many of the non-legal scenarios used in Experiment 1 may have encouraged more lenient thresholds as a false positive would be greatly preferable to a false negative. For example, it may be better to have false positives in determining that: a patient will develop Huntington's disease in order to begin treatment; a stock will under-perform the market in order to avoid losing money; an oil spill has occurred in order to implement clean-up procedures. Experiment 2 sheds additional light on the application of these legal decision standards in a non-legal domain. Indeed, when there was no inherent cost of either a false positive or a false negative we did not observe the same substantial reduction of the decision threshold in the IB condition to below 50%. Chapter 5 of the present thesis further explores the influence of decision consequences by investigating how the potential cost of the decision affects the probability of an affirmative response in different domains.

Finally, we should acknowledge some potential limitations of the current research. First and foremost, our efforts to provide rigorous experimental control and parametric manipulations

favor internal validity at the expense of external validity. For example, multiple pieces of evidence are typically considered together before drawing a legal judgment, and jurors hear from two conflicting points of view. In addition, the difference in decision parameters between the legal and non-legal scenarios in Experiment 1 suggests that scenario content can influence decision-making, possibly due to participants inferring potential outcomes (see above). Future work can explore additional means of presenting multiple (and potentially contradictory) pieces of evidence and/or manipulating the potential outcome associated with a decision while still making use of the psychometric function. Comparing decisions based on reading legal textual scenarios and psychophysical performance in a perceptual task has its own set of challenges and limitations, not the least of which is sacrificing data robustness in the dot motion coherence task of Experiment 3 when attempting to further equate the methodology employed with the legal scenarios of Experiment 1.

These limitations notwithstanding, the upside of comparing burdens of proof instructions and decision standards across domains is substantial. Specifically, the present research applies an experimental approach to examining legal decision-making with psychometric functions, and further compares the legal domain to other domains and to people's intuitive decision standards. By applying a psychometric approach, the findings from this and future work can isolate how variables of interest can influence legal decision making, such as juror instructions, prejudicial information, or punishment outcomes. As such, these studies may yield novel insight into the foundational process of decision-making in law and other societal domains. Understanding how decision criteria are applied in legal and non-legal realms has the potential to instill a more nuanced approach to assessing the factors that impact decision-making by experts and non-experts in a variety of fields.

CHAPTER 3

EFFECT OF EXPERTISE ACROSS INSTRUCTION AND DOMAIN

Introduction

Chapter 2 found that both context and legal burden of proof instructions influence laypeople's decisions, with more stringent decisions in the legal domain versus non-legal domain for all instructions, and markedly distinct decision thresholds associated with the different burdens of proof instructions. Participants' application of the PoE standard was overly stringent in the legal domain compared to the theoretical threshold of 50% – and, surprisingly, compared to people's intuitive belief (IB) – while the BaRD decision threshold was overly lenient in legal and non-legal domains compared to the 90% standard, consistent with previous research (e.g. Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; Simon & Mahan, 1971). In contrast, judges' estimates of the standards are generally comparable to the prescribed thresholds at 50-55% for PoE and around 90% for BaRD (McCauliff, 1982; Simon & Mahan, 1971), indicating that expertise may play a role in the application of these instructions. Previous research has not assessed legal expert decision thresholds in the absence of any instruction to ascertain their intuitive thresholds, or in comparison to other types of expertise. In Chapter 3 we assess expert decisions to determine what their intuitive standards are (for instance, do legal experts adopt a 50% or 90% decision threshold even in the absence of a legal instruction) and to determine the effect of different types of expertise on decision-making and the application of legal instructions.

The legal field is not unique in having adopted arbitrary decision thresholds. This is most evidenced in the scientific and medical fields (Should you reject the null hypothesis?, Do the risks

associated with surgery outweigh the potential benefit to the patient?). Initially introduced by Fisher (1925; as cited in Dahiru, 2008), the scientific field has generally converged to a standard decision threshold of 95%, more formally, alpha=0.05, as the threshold for rejecting the null hypothesis. While there has been a recent push to move away from this standard given that it is largely arbitrary and often misunderstood (e.g. Wagenmakers et al., 2018; Wasserstein, 2019), it still pervades the scientific literature. Alpha represents the probability of making a Type I error by incorrectly rejecting a true null hypothesis (i.e. false positive; Dahiru, 2008). Reducing alpha decreases the risk of a Type I error but can increase the risk of a Type II error in which a true difference is not detected, so researchers typically apply a relatively stringent alpha (0.05 or less) in order to reduce Type I errors while also seeking to maximize power to reduce Type II errors. By comparison, the use of decision thresholds in the medical domain is far more varied due to the vast range of potential diagnoses and treatments. While the 95%/alpha=0.05 threshold is commonly used to report research findings in medical journals, in practice physicians rely on several approaches to making diagnostic and treatment decisions. Models of medical decision-making suggest that physicians often rely on a probabilistic framework to determine the likelihood that a patient has a given diagnosis based in particular on patient symptoms, history, as well as the base rate for each potential diagnosis (Hausmann, Zulian, Battegary, & Zimmerli, 2016; Woolever, 2008). After establishing potential diagnoses physicians typically conduct tests to obtain additional information that can either increase or decrease the likelihood of a given diagnosis, and there is evidence to suggest that they employ a threshold model in which there is some probability at which the physician has enough evidence to make a diagnosis or at which a certain treatment is deemed appropriate (Boland & Lehmann, 2010; Djulbegovic et al., 2014). The treatment threshold varies depending on the associated benefits, risks, and confidence in a given diagnosis. In analogy to the

legal field, these thresholds are not quantified, but instead are often based on experience, confidence, and patient-specific factors (Hausmann et al., 2016). Further, there is substantial variation in physicians' thresholds both for diagnosis and for treatment of the same disease (Boland & Lehmann, 2010; Plasencia, et al., 1992).

This brief review of thresholds across societal domains begs the question of how they relate to one another. Is there a convergence of decision thresholds across societal domains, and if not, what might account for the differences? How do these domain-specific thresholds relate to people's individual beliefs? And, finally, how does one's expertise in a given domain (e.g. a legal expert) affect their standards of proof in other societal domains? These are the central questions that are the main focus in the present study. We address these questions by recruiting cohorts of experts from different professional fields (legal, medical, and scientific, with humanities scholars as education controls and laypeople as baseline controls) and ask them to make decisions in their own field of expertise as well as in other fields using the various burdens of proof (IB, PoE, BaRD) employed in our other studies. As before, we took advantage of the psychometric approach to analyze the data sets.

Methods

Participants

We recruited four distinct cohorts of expert participants and a control group of non-experts. For the control group of non-experts, we recruited 1081 participants (51% male, Mean age=42.50, range=18-76) from the US via Amazon Mechanical Turk. Participants were paid $0.75 for completing the study, which took between five and eight minutes on average. We excluded 23 participants who indicated that they had a PhD, JD, or MD since our goal for the control group

was to recruit participants who did not qualify as an expert in one of our domains of interest (see below).

The expert cohorts consisted of the following: 1) Legal experts (currently working in the legal domain and have a law degree (e.g. JD, LLM), corresponding to faculty from U.S. law schools and currently practicing attorneys, 2) Medical experts (currently working in the medical domain and have a medical degree (e.g. MD, DO), corresponding to faculty from U.S. medical schools, 3) Scientific experts (working in a scientific field and have a PhD in a scientific field), corresponding to faculty from U.S. R1 universities, 4) Humanities experts (currently working in the humanities and have a PhD in a humanities field), corresponding to faculty from U.S. R1 universities. The humanities experts are an education-matched control group with no standard field criteria to serve as a comparison. Expert participants were contacted via email with a description of the study and a link to participate. We contacted 13,348 potential legal experts and had 941 completed responses. After the attention check exclusion we included 764 legal experts in subsequent analyses (61% male; Mean age=49.92, range=27-83). We contacted 34,878 potential medical experts and had 992 completed responses. After the attention check exclusion we included 731 medical experts (54% male, Mean age=48.55, range=28-84). We contacted 20,998 potential scientific experts and had 1498 completed responses. After the attention check exclusion we included 1192 scientific experts (64% male, Mean age=47.89, range=27-89). We contacted 18,521 potential humanities experts and had 1102 completed responses. After the attention check exclusion we included 823 humanities experts (50% male, Mean age=48.11, range=27-81). As compensation, participants had the opportunity to enter a drawing for 10 $50 Amazon gift cards. All participants provided informed consent, and the experimental protocol was approved by the Vanderbilt University Institutional Review Board.

Design and Materials

The task employed a 4 (domain: legal, medical, scientific, control; within-subjects) x 9 (objective evidence strength: 0%, 20%, 40%, 60%, 80%, 90%, 95%, 99%, 100%; within-subjects) x 3 (decision criteria instruction: IB, PoE, BaRD; within-subjects) design. Each participant read and responded to four scenarios, one per domain. The scenarios were presented in random order, followed by a final attention check scenario. The attention check was structured in the same way as the trial scenarios so as to appear to be just another scenario except that it contained a specific instruction within the text that allowed us to identify participants who were not reading scenarios before responding. All of the scenarios used are included in the Supplementary Materials.

The experiment was administered using the Qualtrics online survey platform. The trial design was identical to the trial described in Chapter 2 (Experiment 1, Fig. 1). Evaluation of the scenario/instructions and the subsequent decisions were self-paced.

Participants read one scenario from each domain with random assignment to the scenario within each domain. Scenarios in the **legal domain** were identical to those used in Chapter 2, with three possible fact patterns (stealing prescription drugs, stealing company data, theft, murder) and three types of evidence (video facial recognition, fingerprints, DNA), thus forming nine possible legal scenarios. The objective evidence strength was given as the level of certainty with which investigators concluded that the evidence was left by the protagonist, Mark, presented as a frequentist measure of probability that was randomly assigned from nine possible levels within-subjects: 0%, 20%, 40%, 60%, 80%, 90%, 95%, 99%, 100%.

We also used the same five non-legal scenarios as in Chapter 2 (Huntington's disease, stock underperforming, abnormal water temperatures in the Pacific Ocean, petroleum oil spill, electronic

spam detection) but added additional language to each in order to make them more comparable in length to the legal scenarios. We developed new non-legal scenarios as well in order to have three scenarios in each domain.

Within the **medical domain**, participants saw one of three fact patterns related to a patient's medical test result and potential diagnosis. They judged the likelihood of either: a patient developing Huntington's disease based on DNA markers, a patient having Irritable Bowel Disease (IBD) based on a stool sample, or a patient having optic nerve damage based on their intraocular pressure. The objective evidence strength was again presented as a frequentist probability (same levels as above) for the level of certainty of the patient having the disease given the test result (e.g. When patients are found to have this level of fecal calprotectin, it can be concluded with 80% certainty that the patient has IBD).

The **scientific domain** scenarios consisted of three fact patterns describing the likelihood of either: above average water temperatures developing in the Pacific Ocean based on meteorological patterns, a river being contaminated with petroleum based on water sample analyses, or a near-Earth object colliding with the Earth based on astronomical measurements. The objective evidence strength was the level of certainty for the event having occurred/occurring in the future given the data, and was presented as the same frequentist levels as in the other domains (e.g. When this level of petroleum is detected it can be concluded with 80% certainty that there has been a petroleum spill in the river).

Finally, the **general (miscellaneous) domain** consisted of three fact patterns describing the likelihood of one of the following miscellaneous events: a certain stock underperforming the market based on average price history, an incoming electronic packet containing spam based on a detection system, or a house remaining on the market for six months based on features of the house.

The objective evidence strength was presented in the same manner as described above (e.g. Based on these eight features it can be concluded with 80% certainty that this house will still be on the market in six months).

Participants were randomly assigned to one of the three instruction decision criteria conditions (IB, PoE, BaRD) for each scenario. The language used for the instructions was identical to that of Chapter 2 (Experiment 1, Fig. 2).

As in Chapter 2, participants responded to a yes/no question about their beliefs regarding the scenario. The prompt language was specific to the scenario and instruction criteria (e.g. "Do you believe that Mark stole the prescription drugs?"; "Do you believe by a preponderance of the evidence that the patient will develop Huntington's disease?"). Participants were then asked to provide their own subjective probability for the event by dragging a bar along a 0 to 100 number line. We collected basic demographic information (gender, age, race) as well as information about the degree held by participants in order to confirm that they met our criteria as experts.

Statistical Analyses

We used the same procedures described in detail in Chapter 2 to fit psychometric curves, estimate parameters, and obtain 95% confidence intervals for the parameters using 1000 bootstrap samples. As in Chapter 2, we then generated distributions of the difference scores for our 1000 samples to conduct pairwise comparisons of our parameters to assess differences between conditions by obtaining Bonferroni-corrected confidence intervals for the differences. Confidence intervals that do not contain 0 indicate a significant difference.

Results

First, collapsing the data across decision criteria instruction and domain revealed differences in decision-making based on expertise (Fig. 1; see Table S1-S2 for all parameters and comparisons). Non-expert controls rendered more lenient decisions than all of our expert groups; this was evidenced by a significantly lower decision threshold for non-experts versus all other groups (Non-Expert vs Legal: -3.56[-6.43, -0.76]; Non-Expert vs Medical: -9.99[-12.31, -7.17]; Non-Expert vs Scientific: -10.58[-12.71, -8.14]; Non-Expert vs Humanities: -6.27[-8.80, -3.67]), as well as a greater lower bound for non-experts compared to legal experts (0.03[0.01, 0.07]) and scientific experts (0.03[0.01, 0.05]). Between our expert groups, both legal experts and humanities experts had more lenient decision thresholds than medical (Legal vs Medical: -6.43[-9.65, -3.44]; Humanities vs Medical: 3.72[0.59, 6.66]) and scientific experts (Legal vs Scientific: -7.02[-10.00, -3.81]; Humanities vs Scientific: 4.31[1.33, 6.99]). Interestingly, legal experts were more conservative at the upper bound than medical (-0.05[-0.09, -0.01]) and scientific experts (-0.04[-0.08, -0.002]). The similarity between the medical and scientific experts is noteworthy, as it may suggest that the decisions of medical experts (at least those in our sample) are influenced by their scientific training.

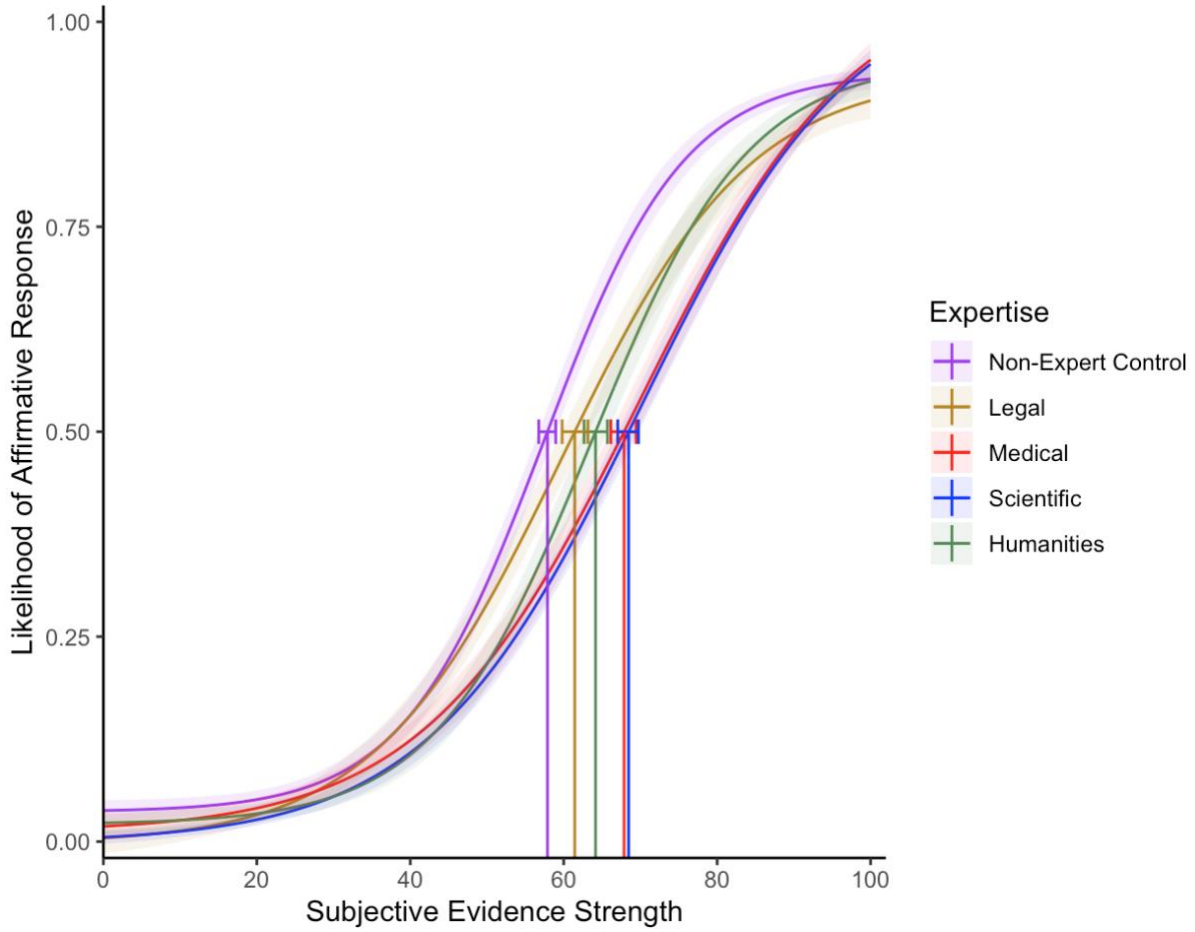

Figure 1. Likelihood of an affirmative response by subjective evidence strength and expertise. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

We then assessed the relationship between expertise and decision criteria instruction by comparing instruction within each expert group (Fig. 2) as well as comparing between expert groups for each instruction (Fig. 3). Figure 4 presents the psychometric parameters for each function (instruction x expertise; see Tables S3-S4 for all parameters and comparisons). The difference and Bonferroni-corrected confidence intervals for the significant comparisons discussed below are presented in Table S4; we do not list them in the text due to the high number of significant comparisons.

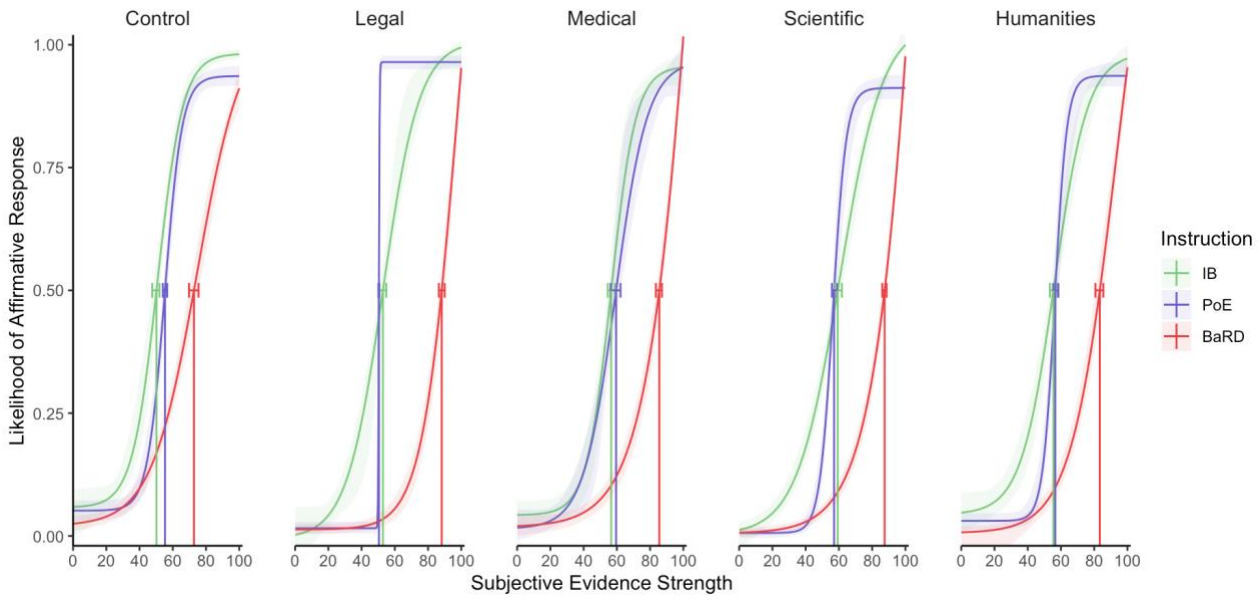

Figure 2. Likelihood of an affirmative response by subjective evidence strength and instruction (colors) and expertise (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

All groups were significantly more stringent for the BaRD instruction than both the IB and PoE instructions, as evidenced by the decision threshold comparisons (Fig. 4A, see also Fig. 2). However, the non-expert controls were the only group that showed a significant difference between IB and PoE, with a greater decision threshold for PoE versus IB, as in the results of Chapter 2. In contrast, the expert groups did not have significantly different decision thresholds between IB and PoE (though interestingly the legal experts and scientific experts had a non-significant lower threshold for PoE compared to IB). Most striking is the differently shaped curves between instruction types for several of our expert groups, which showed a sigmoid shape for IB, more of a step-function for PoE, and an exponential curve for BaRD. This is also reflected in the slope comparisons (Fig. 4B); the PoE slope for legal, scientific, and humanities experts was significantly steeper than both the IB and BaRD slopes, indicating that these participants treated the PoE instruction as a tipping point with a sharp differentiation between "No" and "Yes"

responses. The exponential shape of the expert BaRD curves suggests that they were very conservative until the highest levels of evidence; this is consistent with the significantly more stringent decision thresholds for all expert groups in comparison to the non-expert control group for the BaRD instruction. Between group comparisons for the PoE instruction showed a significantly more lenient decision threshold and significantly steeper slope for the legal experts versus all other groups (Figs. 3, 4B), consistent with the step-function shape that indicates they were remarkably consistent in applying the threshold as a tipping point. For the IB instruction, the non-expert control decision threshold was significantly lower than for medical, scientific, and humanities experts, while the threshold for legal experts was lower than for medical and scientific experts (Figs. 3, 4A).

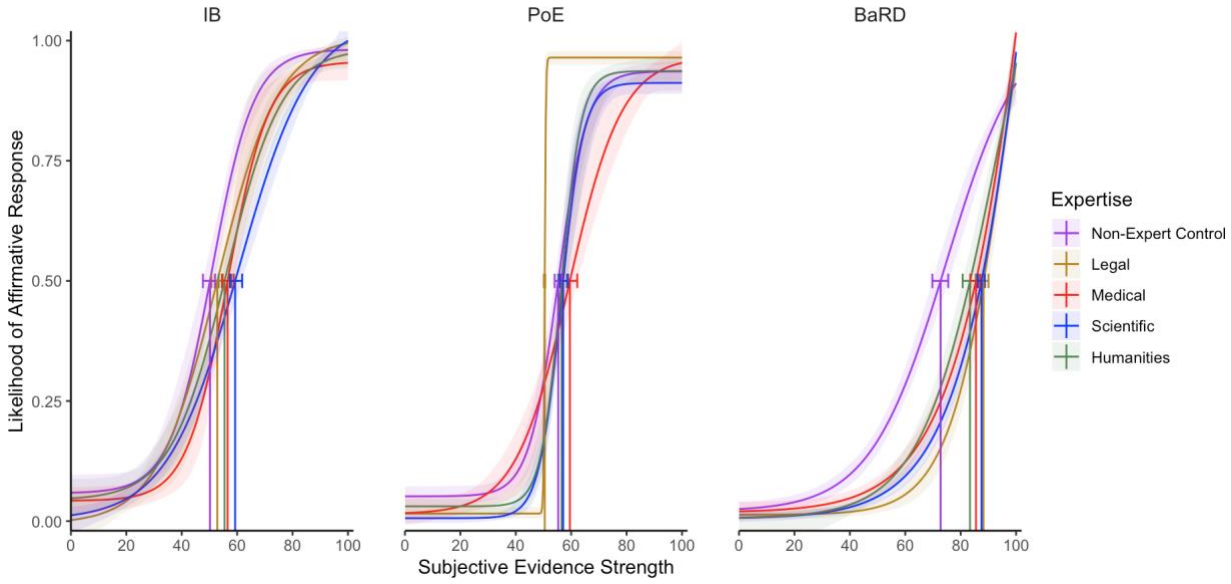

Figure 3. Likelihood of an affirmative response by subjective evidence strength and expertise (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

In addition to comparing decision thresholds between groups, we assessed whether these thresholds were consistent with the prescribed legal standards by determining whether the 95% confidence interval included the 50-55% decision threshold range for PoE (based on the interpretation of this standard by judges; McCauliff, 1982; Simon & Mahan, 1971) or the 90% decision threshold range for BaRD (Fig. 4A, Table S3). PoE thresholds for the medical (59.53 95% CI[56.26, 62.20]), scientific (57.10 95% CI[55.80, 58.87]), and humanities experts (56.58 95% CI[55.30, 58.44]) were greater than the PoE standard, while the threshold for the non-experts (55.33 95% CI[54.02, 56.64]) overlapped with the upper part of the 50-55% range. Remarkably, the threshold for legal experts (50.45 95% CI[50.13, 50.64]) was highly consistent with the theoretical 50% standard. The IB thresholds were consistent with the PoE standard with the exception of the scientific experts who had an IB threshold greater than the 50-55% range (59.27 95% CI[57.42, 61.82]). As for the BaRD decision thresholds, they were lower than the 90% standard for all groups (Non-Experts: 72.72 95% CI[69.78, 75.52]; Medical: 85.52 95% CI[83.47, 87.13]; Scientific: 87.51 95% CI[86.09, 88.70]; Humanities: 83.36 95% CI[80.71, 85.48]) with the notable exception of the legal experts (88.23 95% CI[86.51, 90.04]). Furthermore, comparing the difference between the BaRD and PoE thresholds between expert groups demonstrated that all the expert groups showed greater differentiation than the non-expert controls, while the legal experts showed greater differentiation than all other expert groups (Table 1).

BaRD-PoE	Corrected CI	
Legal vs Medical	[6.13, 16.64]	*
Legal vs Scientific	[4.02, 11.83]	*
Legal vs Humanities	[7.15, 16.22]	*
Legal vs Control	[15.23, 25.75]	*
Medical vs Scientific	[-9.48, 2.23]	
Medical vs Humanities	[-6.39, 6.04]	
Medical vs Control	[2.48, 14.94]	*
Scientific vs Humanities	[-0.72, 8.46]	
Scientific vs Control	[7.30, 18.15]	*
Humanities vs Control	[2.58, 15.08]	*

Table 1. Differentiation between BaRD and PoE decision threshold between all expertise groups

Together these findings indicate that legal experts, in contrast to non-experts, apply the PoE and BaRD instructions in a manner consistent with the prescribed standards. Decisions by the medical, scientific, and humanities experts were generally more conservative than the non-experts and legal experts for the IB and PoE instructions. The scientific and humanities experts showed a steeper slope for PoE than for IB and BaRD instructions, suggesting they treated it as a tipping point, they were overly stringent in their application of the PoE standard.

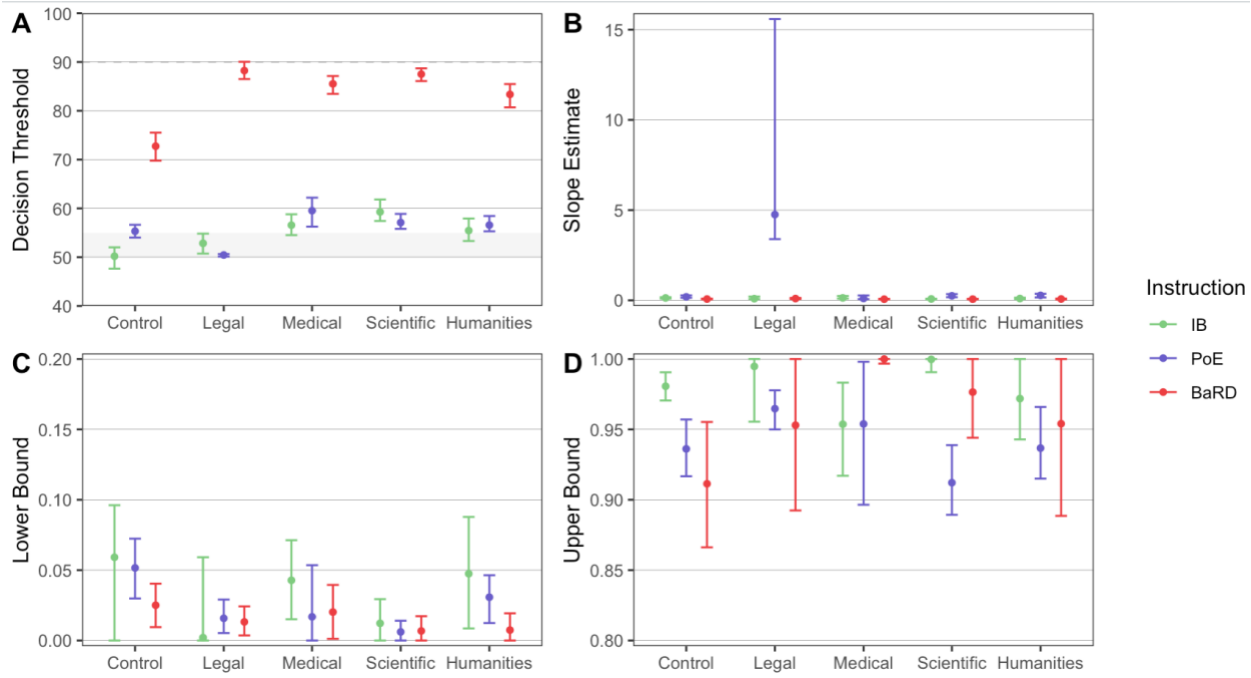

Figure 4. Subjective evidence strength at the decision thresholds (A), slope estimates (B), lower bound (C) and upper bound (D) by instruction type (IB: green; PoE purple; BaRD red) and expertise (x-axes). 95% confidence intervals estimated via 1000 bootstrap samples. The large error bar for the legal PoE slope in B is due to a steeper slope that lead to curves and bootstrap samples having more of a step-function like shape, which resulted in higher slope estimates.

Next we assessed the relationship between domain and expertise by comparing domain within our expert groups (Fig. 5). Notably, the legal and humanities experts did not behave differently between domains. The non-experts were more stringent in the legal domain compared to the medical (6.69[1.20, 12.13]), scientific (6.98[1.61, 11.76]), and general domains (6.07[1.30, 11.08]). Both the medical and scientific experts were more stringent for legal and medical domains versus the scientific and general domains (Medical Experts- Legal vs Scientific: 13.01[6.44, 18.23]; Legal vs General: 12.90[6.22, 20.89]; Medical vs Scientific: 10.21[1.91, 16.43]; Medical vs General: 10.10[0.75, 18.76]; Scientific Experts- Legal vs Scientific: 11.05[4.49, 18.02]; Legal vs General: 16.47[7.05, 20.39]; Medical vs Scientific: 7.36[1.09, 14.82]; Medical vs General: 12.78[2.40, 17.42]). We note that the humanities experts showed a similar but insignificant trend for more stringent decisions for the legal and medical domains as well.

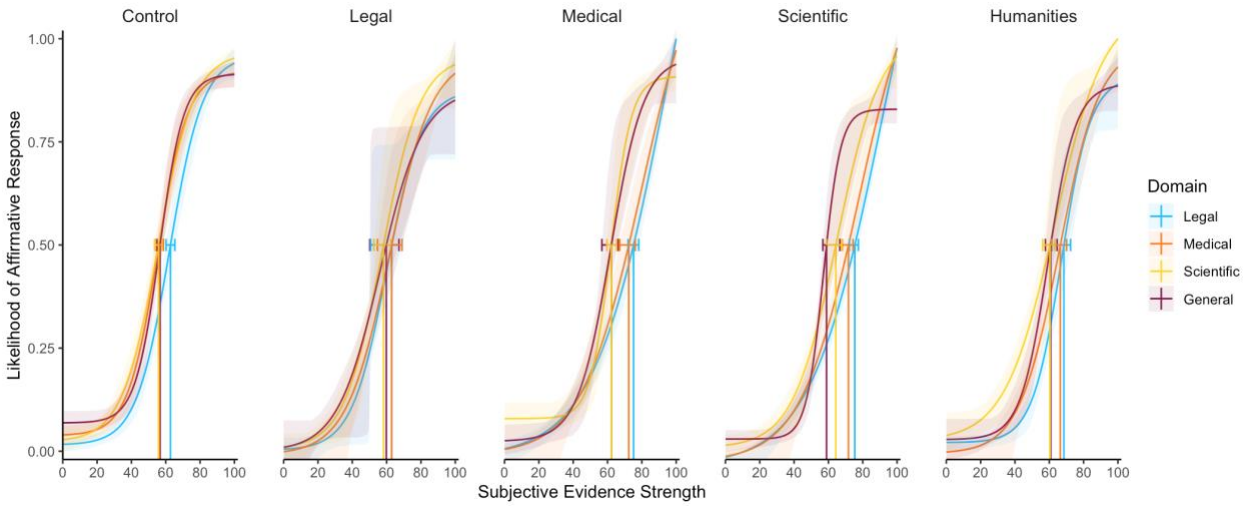

Figure 5. Likelihood of an affirmative response by subjective evidence strength, domain (colors) and expertise group (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

We then compared expertise groups within domain collapsed across instruction (Fig. 6). The medical and scientific experts were more stringent than the other groups in the legal domain (Medical vs Legal: -12.22[-26.64, -2.99]; Medical vs Humanities: 6.67[0.10, 16.47]; Medical vs Controls: -12.42[-18.36, -6.71]; Scientific vs Legal: -12.42[-26.76, -3.50]; Scientific vs Controls: -12.62[-17.12, -6.68]). The expert groups did differ for the medical or scientific domains, and none of the groups differed in decision parameters for the general domain. This suggests that the general domain was a miscellaneous blend of scenarios.

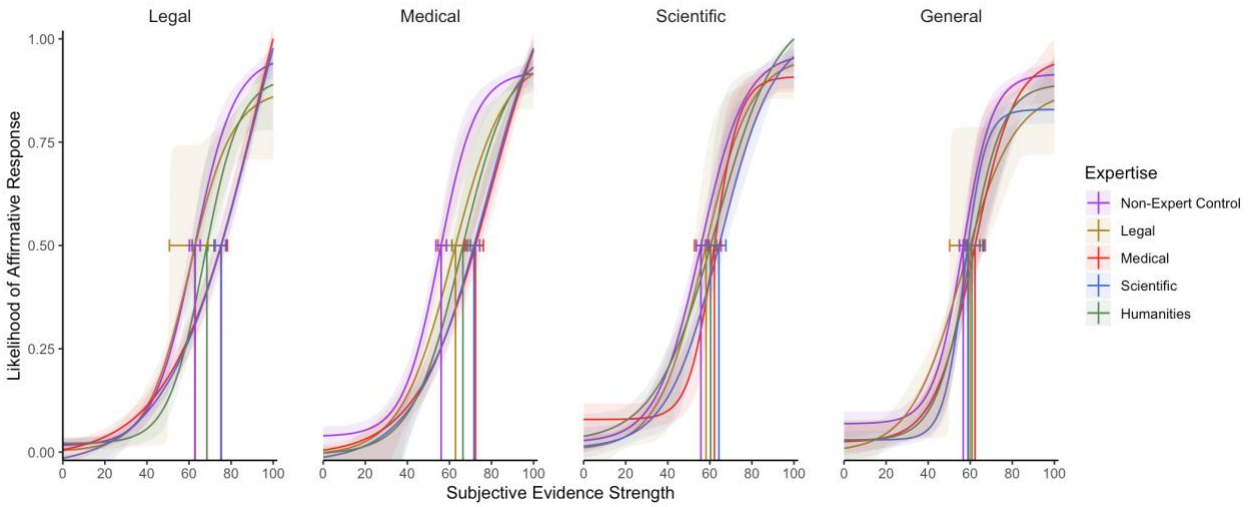

Figure 6. Likelihood of an affirmative response by subjective evidence strength, expertise group (colors) and domain (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Next we assessed the influence of scenario domain by instruction within each expert group. We present Figures 7-11 consecutively to make it easier to compare them. Legal experts showed remarkably little effect of domain, and generally applied the instructions consistently across them (Fig. 7, Tables S5-S6). The only exception was their application of the PoE instruction in the medical domain, which had a more lenient decision threshold than for the legal (5.22[0.64, 7.79]), scientific (-2.19[-3.78, -0.23]), and general domains (-2.53[-4.18, -0.51]), and a less steep slope than the scientific (-11.84[-114.55, -5.68]) and general domains (-2.95[-46.28, -0.87]). The medical domain was also the only threshold to fall below 50% (47.94 95% CI[46.77, 49.51]), suggesting that legal experts were more liberal for decisions in this domain.

In contrast, medical experts demonstrated several effects of scenario domain (Fig. 8, Tables S7-S8). Most notably, the PoE instruction in the general domain had a steeper slope than the legal (-18.34[-88.38, -2.66]), medical (-18.36[-88.36, -2.68]), and scientific domains (-18.31[-88.27, -1.68]), and a lower decision threshold than the medical domain (10.90[1.86, 17.92]). For this condition participants applied a tipping point of just above 50% (51.00 95% CI[50.49, 51.51]).

The BaRD decision threshold for the legal domain was also significantly greater than the scientific domain (8.00[0.31, 22.18]).

Scientific experts were more conservative for the legal domain (Fig. 9, Tables S9-S10), as seen in a greater BaRD decision threshold for the legal versus scientific domain (5.47[0.15, 11.63]) and a greater IB decision threshold for the legal versus scientific (12.45[3.60, 20.48]) and general domains (11.54[2.76, 18.50]). There was also a noticeable upwards shift for the legal PoE curve, though due to higher variance in responses this was not significantly different from the other domains.

Humanities experts (Fig. 10, Tables S11-S12) were more conservative for IB decisions in the legal domain than the scientific (14.47[4.51, 26.73]) and general domains (13.90[3.74, 23.03]), though additional participants are needed to obtain a wider spread of decisions across subjective evidence strength and to reduce variance, in order to interpret these findings further.

Finally, non-experts were more stringent in their decisions for the PoE instruction in the legal domain (Fig. 11, Tables S13-S14); the decision threshold for the PoE legal domain was greater than the medical (7.37[1.16, 13.87]) and scientific domains (7.22[1.43, 12.52]). Within the legal domain, participants did not distinguish between the PoE and BaRD thresholds (i.e. no difference decision threshold) as they did in the non-legal domains (-14.28[-21.97, 1.89]). The PoE threshold for the legal domain was also greater than the 50-55% prescribed standard (61.01 95% CI[57.30, 64.22]).

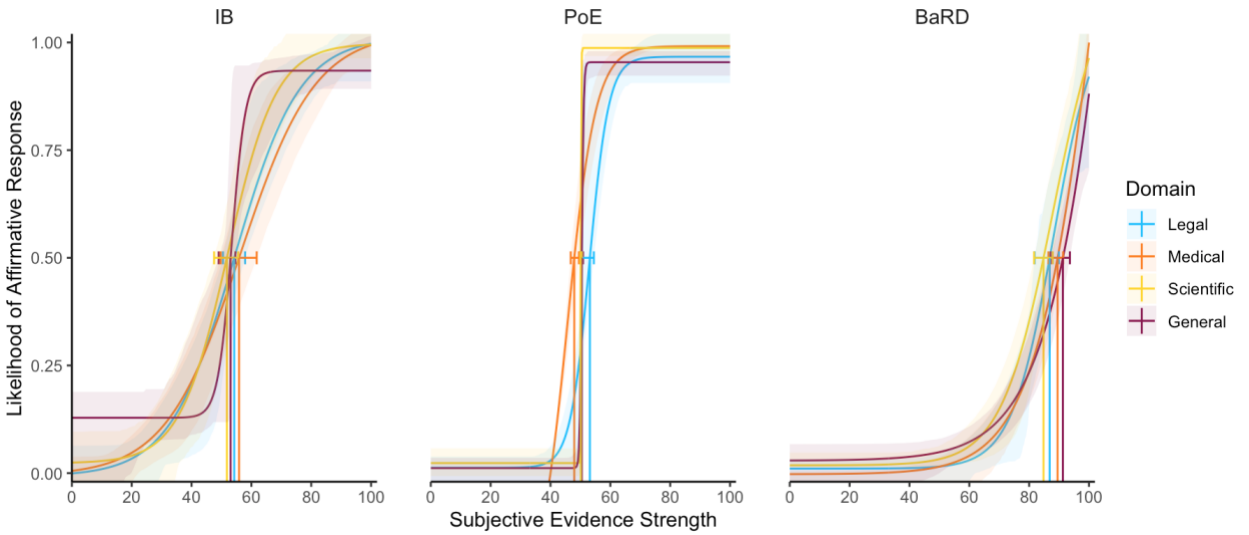

Figure 7. Likelihood of an affirmative response for legal experts by subjective evidence strength, domain (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

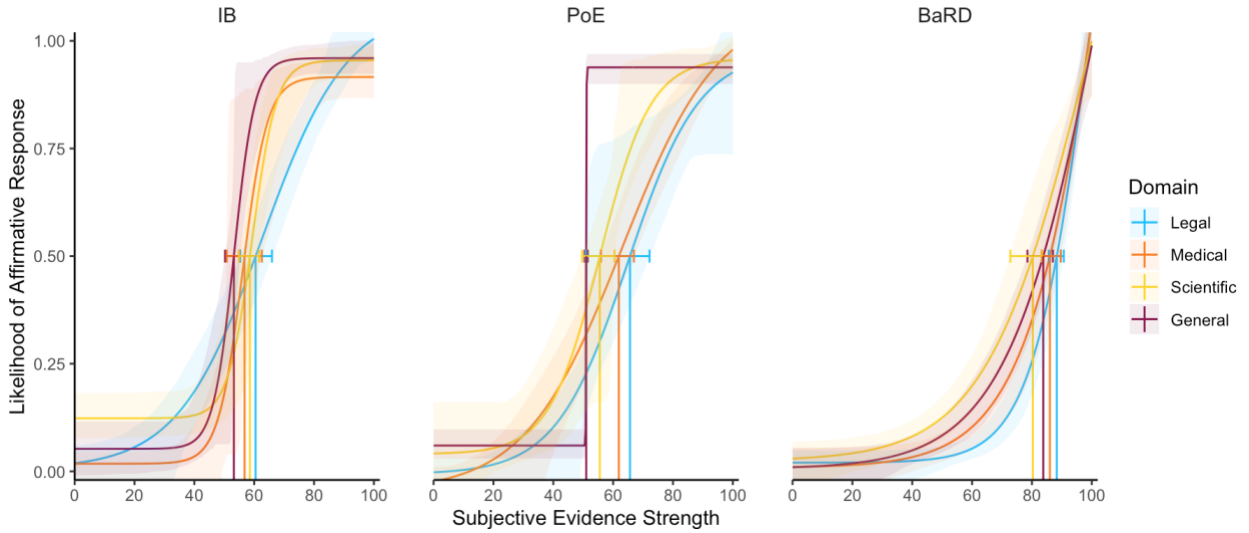

Figure 8. Likelihood of an affirmative response for medical experts by subjective evidence strength, domain (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

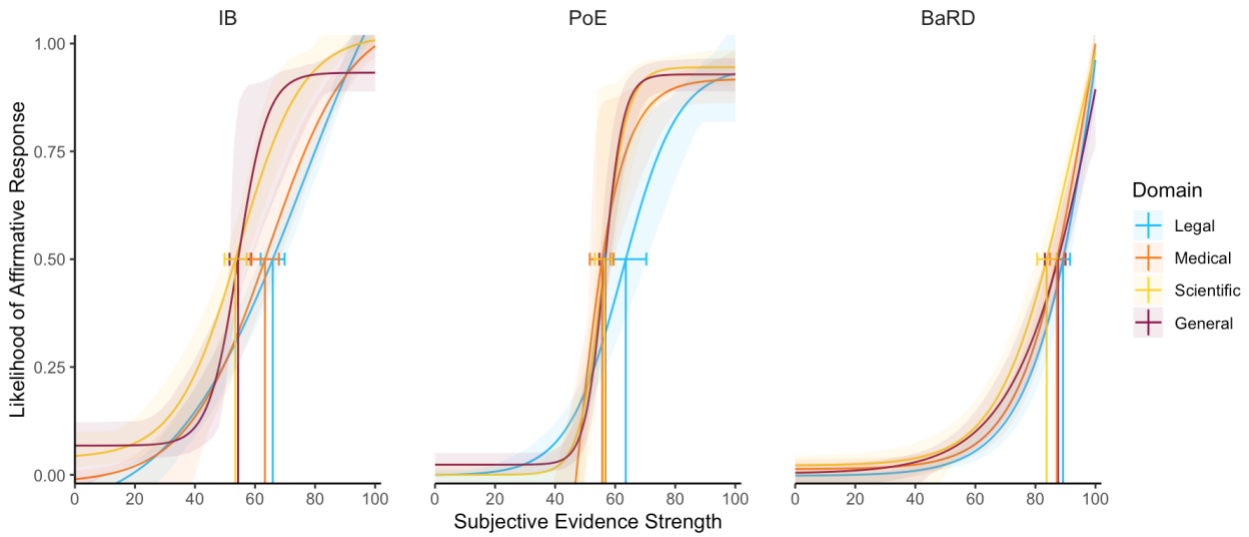

Figure 9. Likelihood of an affirmative response for scientific experts by subjective evidence strength, domain (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

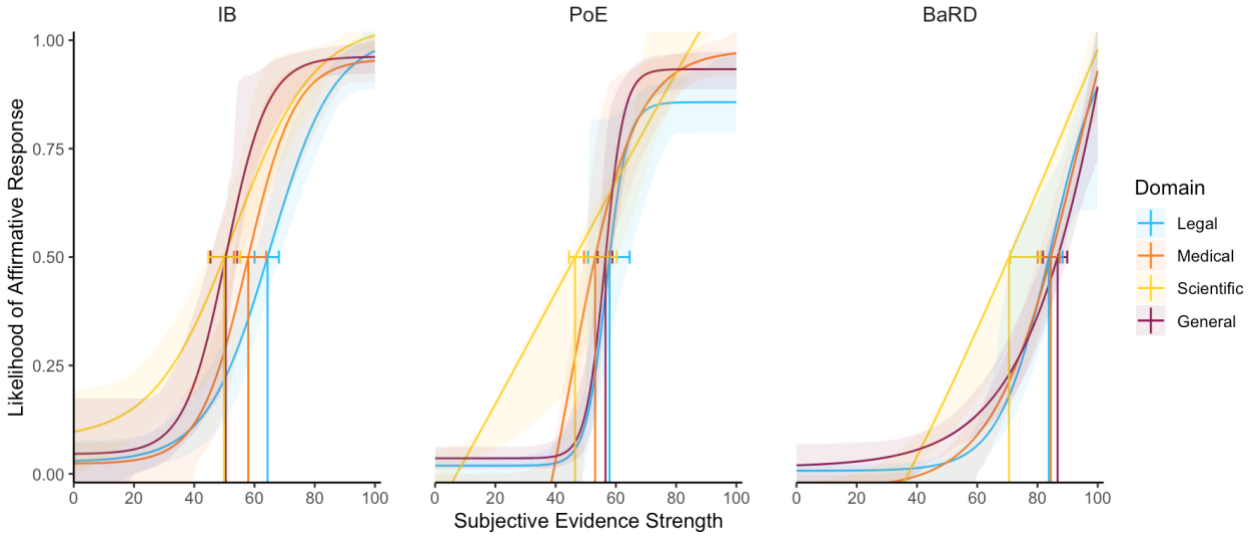

Figure 10. Likelihood of an affirmative response for humanities experts by subjective evidence strength, domain (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

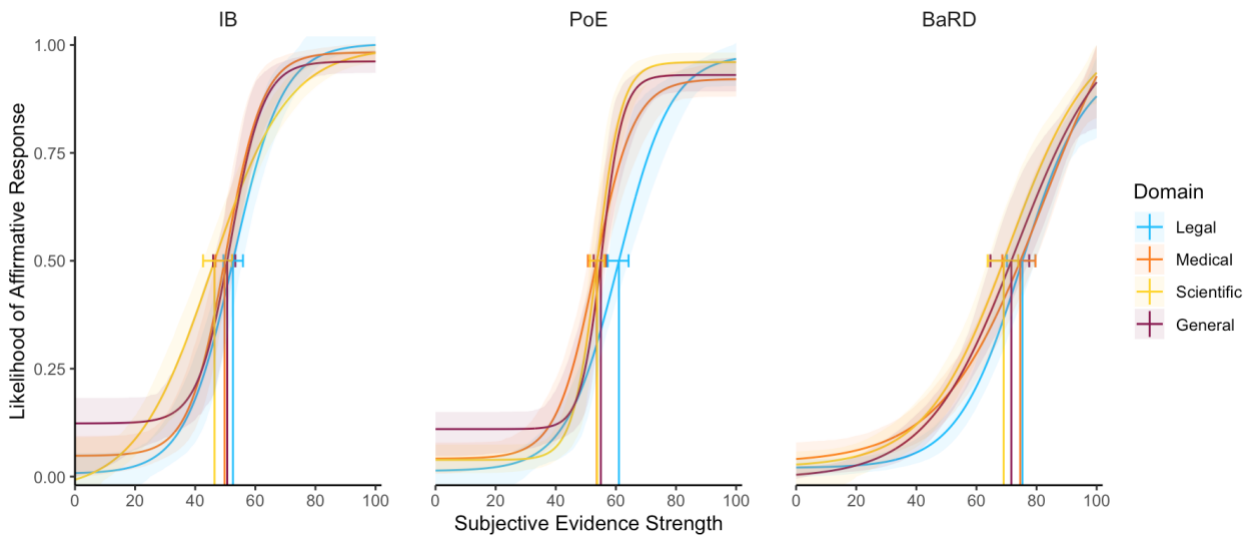

Figure 11. Likelihood of an affirmative response for non-expert controls by subjective evidence strength, domain (colors) and instruction (plot panels). Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Discussion

In this study we assessed the effect of domain of expertise on the application of legal standard instructions across scenario domains. We found that expertise influenced decisions differently depending on both the instruction and domain, as discussed below.

Consistent with our expectations and past research (e.g. McCauliff, 1982; Simon & Mahan, 1971), legal experts applied the PoE and BaRD standards in a manner consistent with the prescribed thresholds of 50% and 90% respectively. This was apparent in the shape of the curves, which were a step-function for PoE instructions and an exponential curve for BaRD. We further evaluated these decisions in a non-instructed condition (IB) and across different societal domains, something not previously examined, and found that legal experts were generally consistent in how they applied each instruction regardless of domain, consistent with the intended use of the PoE and BaRD standard to provide equitable decision thresholds across different cases. The one exception was more liberal decisions for the PoE instruction in the medical domain, which may suggest that participants favored a false positive, possibly because a diagnosis could lead to

treatment. Interestingly, legal experts' intuitive decision thresholds were comparable to their application of the PoE instruction; this indicates that they may have adopted the just over 50% threshold as a go-to means of making decisions. However, the shape of the fits was different which suggests that they may apply the same threshold but not identical behavior. Finally, an important distinctive feature of the legal experts is that they appeared to be more discriminatory in their application of decision standards not only in the legal field, but in all domains (see Fig. 2). This is a clear indication that expertise in the legal domain sharpened individuals' application of the legal standards.

While the other experts may not have been as discriminatory as legal experts in their applications of different burdens of proof, they appeared to have been more conservative in their decisions than non-experts, particularly with regards to their BaRD and IB thresholds (Fig. 3). This was especially true of the scientific and medical experts when collapsing across instruction and domain, where they demonstrated more conservative decisions in comparison to the humanities experts as well. The scientific field generally applies a criterion of 95% or $p<.05$ (sometimes even 99% or $p<.01$) in order to adjudicate among hypotheses (Dahiru, 2008); the increased stringency of these participants could be due to their familiarity with this more conservative decision threshold. That said, it is interesting to note that while science experts took as conservative a stance with BaRD as the legal experts (see Fig. 4A), their decision thresholds were nevertheless well short of the Fisher's 95% hypothesis rejection standard, suggesting that the BaRD threshold, for all of its strict language, is a criterion that is more liberal than the scientific standard. As for the medical experts, it may not be surprising that they showed similar behavior to the scientific experts as they are typically exposed to a substantial amount of scientific training. Furthermore, our sample

was collected from faculty at US medical schools and these individuals may be more likely to be engaged in research than medical experts not affiliated with a university.

Finally, for all but one cohort, we observed an effect of scenario domain such that decisions were generally more conservative in the legal domain. For the non-experts this effect was primarily seen for the PoE instruction, while our non-legal experts demonstrated this effect for the BaRD and IB instructions as well. This is consistent with our finding in Chapter 2 (Experiment 1) that a legal context evoked more stringent decisions. The only exception to this finding is – ironically – the cohort of legal experts, who applied the legal (and IB) standards systematically across all domains. In this case it appears that familiarity/expertise with the legal language dominated the legal experts decision-making irrespective of the domain it is applied to.

We should acknowledge limitations of the present study. In particular, it will be necessary to collect additional data for several of our groups (e.g. humanities) in order to have sufficient data to fully assess the effect of instruction x domain within each group. The fact that we recruited our expert participants from universities may have implications for our findings as well (as noted above for medical experts), and future work could expand to include different expert populations (e.g. practicing attorneys not affiliated with higher learning institutions).

These limitations notwithstanding, the present findings allow us to make general inferences about the effect of expertise in decision making. For one, individuals' expertise to decision standards appear to spill over to other domains than theirs of expertise. This is most markedly observed with legal experts, who not only applied the PoE and BaRD instructions correctly, they used a decision threshold consistent with PoE even in the absence of any instruction, which may indicate that expertise and familiarity with a standard decision threshold influences intuitive decision-making. The relative conservativeness of the scientific and medical experts could also be

consistent with this idea, as these experts have had frequent exposure to the 95% decision threshold associated with hypothesis testing. Relatedly, there is an effect of expertise on decision criterion stringency. Science and medical experts generally applied the most conservative criteria, followed by humanities experts and finally by non-experts (the legal experts were somewhat distinct in their discriminatory application of different decision standards). This may have implications in jury selection, as our data suggest that academics may generally adopt more stringent decision criteria than the general population (as represented by our control group).

Our findings also suggest that there is no such thing as a common or 'universal' intuitive belief decision threshold maintained across individuals; we found about a 10% difference in decision thresholds for IB across cohorts (see Fig. 4A). Such spread could be even larger at the individual level or for other cohorts not tested in the present study. Instead, as mentioned earlier, there is strong evidence that people's domains of expertise influence how they make decision in other domains than those of their expertise.

Finally, the conservative decision stance with legal scenarios is pervasive, applying not only to laypeople (see the present results and Chapter 2) but also to non-legal experts. Thus, no matter the education level or professional domain of expertise, legal scenarios are treated differently than non-legal ones, a point we return to in Chapter 5. One exception to this rule may be the difference in application of the IB and PoE standards between experts and non-experts. While the former did not practically distinguish between IB and PoE, the non-experts did in both the present and in the previous study (Chapter 2). In Chapter 4 we seek to understand why PoE is interpreted more conservatively than both the prescribed threshold as well as IB for non-experts.

CHAPTER 4

APPLICATION OF POE STANDARD ACROSS DOMAINS

Introduction

In Chapters 2 and 3, we assessed how instructions to apply the PoE and BaRD legal burdens of proof influenced individuals' decisions in comparison to when they received no instruction (intuitive belief: IB), across legal, non-legal, and – for Chapter 2 – psychophysical domains. Consistent with previous research (e.g. Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; McCauliff, 1982; Simon & Mahan, 1971), we found that for laypeople (non-experts) the interpretation of the PoE instruction was overly stringent in comparison to the theoretical 50% PoE standard and the 50-55% standard employed by legal experts (McCauliff, 1982; Simon & Mahan, 1971), while interpretation of the BaRD instruction was overly lenient when compared to its theoretical 90% standard. It has been suggested that the discrepancies in laypeople's interpretations of the PoE and BaRD standards are due in part to individuals adjusting each legal standard to better align with their own intuitive decision threshold that falls somewhere between (Arkes & Mellers, 2002; Simon & Mahan, 1971); indeed, interpretations of the BaRD standard were found to be lower for individuals who chose to convict compared to those who acquitted (Park et al., 2016), consistent with the notion that individuals may base their interpretation of the burdens of proof on their own decision threshold. But while the literature previewed our PoE and BaRD results, it did not predict that comparison of these standards with IB within the legal domain would reveal the PoE instruction to beget more stringent decisions than no instruction. Furthermore, within both the legal domain (Chapter 2 Experiment 1, Chapter 3 non-

experts) and the psychophysical domain (Chapter 2 Experiment 2) this difference between IB and PoE was due to participants being overly stringent in interpreting the PoE instruction, as the PoE thresholds were significantly greater than the 50-55% standard while the IB threshold fell closer to 50% (IB not significantly different from 50% for the psychophysical task or Chapter 3). Together these findings suggest that the overly stringent interpretation of PoE cannot be explained by individuals adjusting the standard towards their intuitive decision threshold. If the IB hypothesis cannot account for the real-world interpretation of PoE, what does? The principal aim of this study is to understand why PoE is interpreted as not only more conservative than the prescribed threshold, but also as more stringent than individuals' own belief.

The most likely account for the discrepancies in the interpretation of PoE and BaRD legal standards is the confusion about the meaning of these standards. Most of this work has examined laypeople's understanding of BaRD and has found that jurors report confusion about the term, and that different definitions of BaRD lead to different verdict rates, even when interpretations of the evidence remain consistent (e.g. Cruse & Browne, 1987; Dhami, Lundrigan, & Mueller-Johnson, 2015; Horowitz & Kirkpatrick, 1996; Kerr et al. 1976). In regards to PoE, Kagehiro and Stanton (1985) compared verdict decisions between PoE and BaRD instructions/definitions for the same trial information, and found that participants did not differentiate between the two standards. When the burdens of proof were instead defined using a quantified percentage (51% and 91%), participants differentiated between PoE and BaRD in a manner consistent with the law (i.e. more conservative decisions/ fewer verdicts in BaRD versus PoE condition). This suggests that difficulty in understanding what the PoE standard means may contribute to discrepancies in the way it is applied by laypeople, although the above study assessed verdicts for PoE only relative to BaRD and did not obtain an estimate of participants' threshold for either. They also compared several

different sets of PoE and BaRD definitions and found that this influenced participants' ability to differentiate between these standards, although conclusions could only be made about each set as a whole since the PoE and BaRD definitions within a given set were always presented together (Kagehiro & Stanton, 1985). One additional explanation that has been put forth is that confusion and unfamiliarity with the PoE instruction leads laypeople to apply a threshold more akin to BaRD, which they have heard used in media (i.e. movies, television; Winter, 1971 as cited in McCauliff 1982). Interestingly, judges were found to rank the phrase "preponderance of the evidence" as a more stringent standard than its theoretical equivalent (and definition) "more probable than not" (McCaulliff, 1982), suggesting that even among legal experts PoE may be interpreted differently than the theoretical threshold of just over 50%. Furthermore, judges' interpretations of the value of PoE were found to fall between 50-55% on average (McCauliff, 1982; Simon & Mahan, 1971). Because PoE is likely to be unfamiliar legalese for most laypeople, it would make sense that the discrepancy between 'preponderance of the evidence' and analogous phrases may be more pronounced for laypeople than the small difference seen for judges, not only due to confusion regarding what it means but also because it may imply a legal context. While past research provides evidence that the specific wording of the PoE instruction may influence how it is applied, we note that in Chapters 2 and 3 the PoE threshold for the non-legal context was not significantly different from the prescribed standard, in contrast to our findings with the legal scenarios (even though the instruction/definition was identical between legal and non-legal context), suggesting that the instruction alone does not provide the whole story and that its interpretation may be context-dependent. Indeed, a thorough account for PoE's effects on legal decision-making must not only explain the variance associated with different definitions of this legal standard, but also why it is interpreted more stringently than its prescribed threshold within a legal context.

The present study seeks to better understand the relationship between the PoE standard and context by comparing multiple instructions that are all theoretically equivalent to a decision threshold of just over 50%, across legal and non-legal societal domains, using a psychometric approach. As described in Chapter 2, a psychometric approach is advantageous as it allows us to compare and quantify decisions across a range of evidence strengths, and the four psychometric parameters each provide insight into different aspects of the decision-making process that together yield to a thorough understanding of the effects of our variables. We assessed decisions across the same domains as in Chapter 3, namely legal, medical, scientific, and general (miscellaneous). Within each domain we compared three PoE instructions (two with definitions and one without) as well as a quantified instruction (>50%), "more likely true than not" (which itself is a definition of PoE), and IB. We expected that the PoE instructions would result in more stringent decisions than the other conditions across domain due to the inclusion of a less familiar legalese term, and we sought to determine whether different definitions or a lack thereof affected the interpretation of PoE. Based on the relationship between instruction and domain in Chapters 2 and 3 we further hypothesized that the most stringent decisions would be for the PoE instructions within the legal domain. If these hypotheses are supported it would suggest that the phrase PoE itself leads to more stringent decisions, and that this effect is amplified within the legal domain. Just as importantly, by comparing the effect of these different definitions to the prescribed decision criterion and to IB, we hoped to identify the one(s) that may serve best the intent of the justice system.

Methods

Participants

We recruited 4439 participants from the United States via Amazon Mechanical Turk. Of these, 1570 were eliminated for failing the attention check (see below) leaving a total of 2869 participants 1420 male 10 other (50% male; Mean age= 38.60 years, Range= 18 to 84 years) for all subsequent analyses. Participants were paid $0.75 for completing the task, which took between five and eight minutes on average. All participants provided informed consent and the experimental protocol was approved by the Vanderbilt University Institutional Review Board. Consistent with the previous Chapters, we recruited participants until we reached roughly 40 observations per cell after exclusions. The study was preregistered using the Open Science Framework (https://doi.org/10.17605/OSF.IO/FZ6MJ).

Design and Materials

The task employed a 4 (domain: legal, medical, scientific, general; within-subjects) x 7 (objective evidence strength: 0%, 20%, 40%, 60%, 80%, 90%, 100%; within-subjects) x 6 (decision criteria instruction: see below; between-subjects) design. Each participant responded to four trial scenarios, one per domain. The scenarios were presented in random order, followed by a final attention check scenario. The attention check was structured in the same way as the trial scenarios so as to appear to be just another scenario except that it contained a specific instruction within the text that allowed us to identify participants who were not reading scenarios before responding. All of the scenarios used are included in the Supplementary Materials.

After recruitment via Mechanical Turk, participants were directed to the Qualtrics online survey platform to complete the experiment. The trial design was identical to the trial described in Chapter 2 (Experiment 1, Fig 1). Participants read the scenario/instructions and responded at their own pace.

Participants were randomly assigned to one scenario in each of our four domains (legal, medical, scientific, general), presented in random order. Scenarios in the **legal domain** described criminal or civil wrong-doing as well as evidence linking a protagonist, Mark, to the act. As in our previous studies (Chapters 2 and 3), we had nine possible legal scenarios stemming from three fact patterns (stealing prescription drugs, stealing company data, and murder) crossed with three types of evidence (video facial recognition, fingerprints, and DNA). The objective evidence strength was given as the level of certainty with which investigators concluded that the evidence was left by the protagonist, presented as a frequentist measure of probability that was randomly assigned from seven possible levels within-subjects: 0%, 20%, 40%, 60%, 80%, 90%, 100%. As in the previous experiments there was no significant difference between fact patterns or evidence so we collapsed across both (Fig. S1-S2, Tables S1-S2).

The non-legal scenarios were identical to those used in Chapter 2. Within the **medical domain**, participants saw one of three fact patterns related to a patient's medical test result and potential diagnosis. They judged the likelihood of either: a patient developing Huntington's disease based on DNA markers, a patient having Irritable Bowel Disease (IBD) based on a stool sample, or a patient having optic nerve damage based on their intraocular pressure. The objective evidence strength was again presented as a frequentist probability (same levels as above) for the level of certainty of the patient having the disease given the test result (e.g. When patients are found to have this level of fecal calprotectin, it can be concluded with 80% certainty that the patient has IBD). As in Chapter 2 there was not a significant difference between fact patterns so we collapsed across them (Fig. S3, Table S3).

The **scientific domain** scenarios consisted of three fact patterns describing the likelihood of either: above average water temperatures developing in the Pacific Ocean based on

meteorological patterns, a river being contaminated with petroleum based on water sample analyses, or a near-Earth object colliding with the Earth based on astronomical measurements. The objective evidence strength was the level of certainty for the event having occurred/occurring in the future given the data, and was presented as the same frequentist levels as in the other domains (e.g. When this level of petroleum is detected it can be concluded with 80% certainty that there has been a petroleum spill in the river). While we observed more lenient decisions for the chemical spill fact pattern, we collapsed across fact patterns for our primary analyses (Fig. S4, Table S4).

Finally, the **general (miscellaneous) domain** consisted of three fact patterns describing the likelihood of either: a certain stock underperforming the market based on average price history, an incoming electronic packet containing spam based on a detection system, or a house remaining on the market for six months based on features of the house. The objective evidence strength was presented in the same manner as described above (e.g. Based on these eight features it can be concluded with 80% certainty that this house will still be on the market in six months). As in Chapter 3 there was not a significant difference between fact patterns so we collapsed across them (Fig. S5, Table S5).

Participants were randomly assigned to one of six **decision criteria instructions** that remained the same for each of the scenarios they evaluated (i.e. manipulated between subjects). These instruction types are described below with a sample of the language used for each, and for a rationale for their inclusion in this experiment when appropriate.

1). <u>"Intuitive belief" (IB)</u> - no instruction provided in order to assess participants' intuitive decisions. Participants made their decision immediately after reading the scenario.

2). <u>PoE w/ Definition</u> - instructed participants to apply the PoE standard and provided a definition adapted from existing jury instructions (U.S. District Court N.D. Cal., 2012);

this is the PoE instruction we have used previously and in related studies (Chapters 2, 3, 5). This also served as a comparison to our previous experiments.

Start from a presumption that Mark did not steal the drugs. This presumption requires you to conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark stole the drugs by a preponderance of the evidence. That means that the evidence produced leads you to believe that Mark having stolen the drugs is more likely true than not. To put it differently, if you were to put the evidence favoring Mark having stolen the drugs on one side of a balance scale and the evidence favoring Mark not having stolen the drugs on the opposite side, the evidence has to make the scale tip somewhat in order to conclude that Mark stole the prescription drugs.

3). PoE w/ Alternate Definition - instructed participants to apply the PoE standard and provided a different definition (of similar length to above) of the standard adapted from existing jury instructions (California Jury Instructions- BAJI 2.60). We included this condition in order to determine whether the definition associated with a PoE instruction influences the interpretation of the standard.

Start from a presumption that Mark did not steal the drugs. This presumption requires you to conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark stole the drugs by a preponderance of the evidence. That means that the evidence produced favoring Mark having stolen the drugs has more convincing force than that opposed to it. If the evidence is so evenly balanced that you are unable to say that the evidence favoring Mark having stolen the drugs has more convincing force than the evidence favoring Mark not having stolen the drugs, you must conclude that he did not steal the drugs

4). PoE w/ No Definition - instructed participants to apply the PoE standard with no additional definition given. We included this condition to serve as a control in comparison to the two PoE instructions with definitions, in order to determine whether definitions improve (or possibly hinder) understanding of PoE.

Start from a presumption that Mark did not steal the drugs. You must conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark stole the drugs by a preponderance of the evidence.

5). "More Likely True Than Not" - instructed participants to apply "more likely true than not" to their decision- this is a theoretical equivalent and the definition of "preponderance of the evidence". This was included to determine whether participants can apply a threshold consistent with PoE when they receive an instruction that does not contain legalese.

> Start from a presumption that Mark did not steal the drugs. You must conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that Mark having stolen the drugs is more likely true than not.

6). Quantified - instructed participants to make their decision based on whether the evidence is greater than 50% - this is another theoretical equivalent of "preponderance of the evidence". While Kagehiro and Stanton (1985) found that quantified instructions improved participants' ability to distinguish between PoE and BaRD, they did not obtain estimates of the threshold applied to these standards, so we have included a quantified for the present experiment.

> Start from a presumption that Mark did not steal the drugs. You must conclude that Mark did not steal the drugs unless you are satisfied that the facts above proved that the likelihood that Mark stole the drugs is greater than 50%.

Participants made a "Yes" or "No" response for each scenario in response to a question about their beliefs regarding the action or event described (e.g. "Do you believe that Mark stole the prescription drugs?"; "Do you believe that the patient will develop Huntington's disease?"; "Do you believe that there has been a petroleum spill in the river?"; "Do you believe that the house will still be on the market in six months?"). Following this decision, participants proceeded to a new screen that asked them to give their own subjective estimate of the probability that the action or event had or would happen (e.g. "What do you believe is the probability that Mark stole the prescription drugs?"; "What do you believe is the probability that the house will still be on the market in six months?"), as in Chapter 2.

After completing the four trial scenarios, participants responded to an attention check scenario (that they believed was part of the main task). This scenario was identical to the trial scenarios in its design and layout but contained specific language to provide a specific response on the subsequent screen (see Supplementary Materials). Those who passed the attention check and successfully completed the survey then provided demographic information and were debriefed.

Statistical Analyses

We used the same procedures described in detail in Chapter 2 to fit psychometric curves, estimate parameters, and obtain 95% confidence intervals for the parameters using 1000 bootstrap samples. As in Chapter 2, we then generated distributions of the difference scores for our 1000 samples to conduct pairwise comparisons of our parameters to assess differences by instruction type within each domain (15 comparisons to compare between all instruction types; Bonferroni-corrected CI=1-(0.05/15)) as well as to assess differences by domain within instruction type (6 comparisons to compare between all domains; Bonferroni-corrected CI=1-(0.05/6)).

Results

Collapsing across instruction type, as hypothesized and consistent with Chapter 3 we observed a clear effect of domain such that decisions were most stringent in the legal domain as compared to any of the non-legal domains (Fig. 1). Specifically, participants required stronger evidence in order to make an affirmative response in the legal domain as evidenced by the significantly greater decision threshold for that domain compared to all other domains (Legal vs

Medical: 4.36[2.06, 6.76]; Legal vs Scientific: 4.71[2.07, 7.17]; Legal vs General: 4.71[2.07, 7.17]; see Tables S6-S7 for all parameter estimates and comparisons).

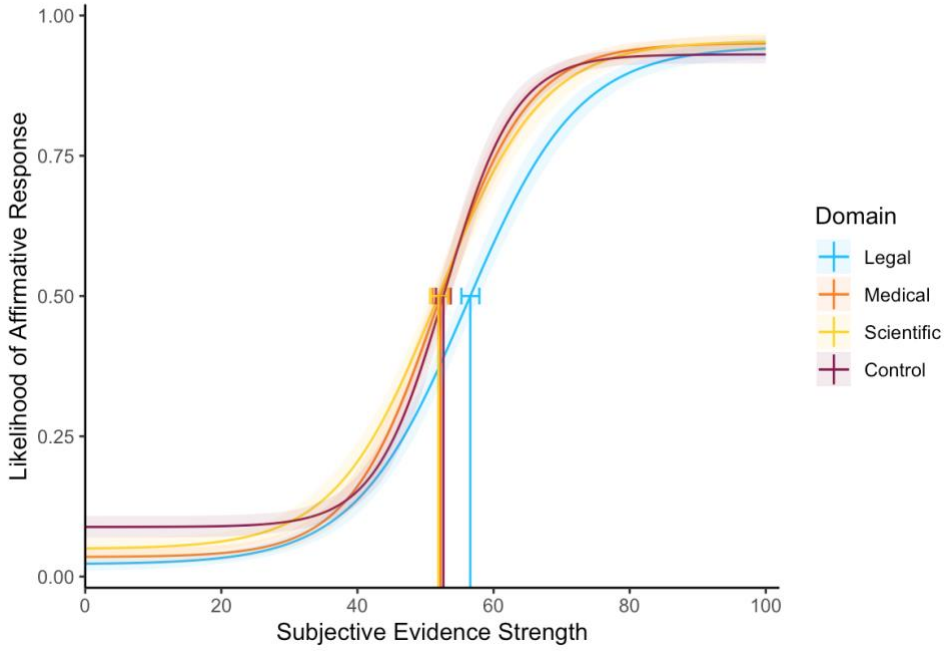

Figure 1. Likelihood of an affirmative response by subjective evidence strength and domain. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

We then assessed the effect of instruction collapsed across domain (Fig. 2; see Tables S8 and S9 for all parameter estimates and comparisons). Participants were significantly more lenient when they received no instruction (IB) than when they received any sort of instruction, as indicated by comparisons of the decision thresholds (PoE w/ Definition vs IB: 5.71[2.22, 8.73]; PoE w/ Alt Definition vs IB: 9.27[5.19, 12.66]; PoE w/ No Definition vs IB: 8.66[4.96, 12.75]; "More Likely True Than Not" vs IB: 5.38[2.17, 8.72]; Quantified vs IB: 6.03[3.41, 8.81]). This appears to be due primarily to participants being particularly liberal in this condition as the IB threshold was lower than 50% (47.30 95% CI[45.64, 48.85]), which indicates that they may have favored false positives in the absence of any instruction. Our hypothesis that the PoE instructions would be most

stringent was partially supported, as the highest decision thresholds were for the PoE w/ Alt Definition and PoE w/ No Definition instructions and both were significantly greater than the "More Likely True Than Not" instruction (PoE w/ Alt Definition vs "More Likely True Than Not": 3.89[0.17, 7.35]; PoE w/ No Definition vs "More Likely True Than Not": 3.28[0.02, 6.58]).

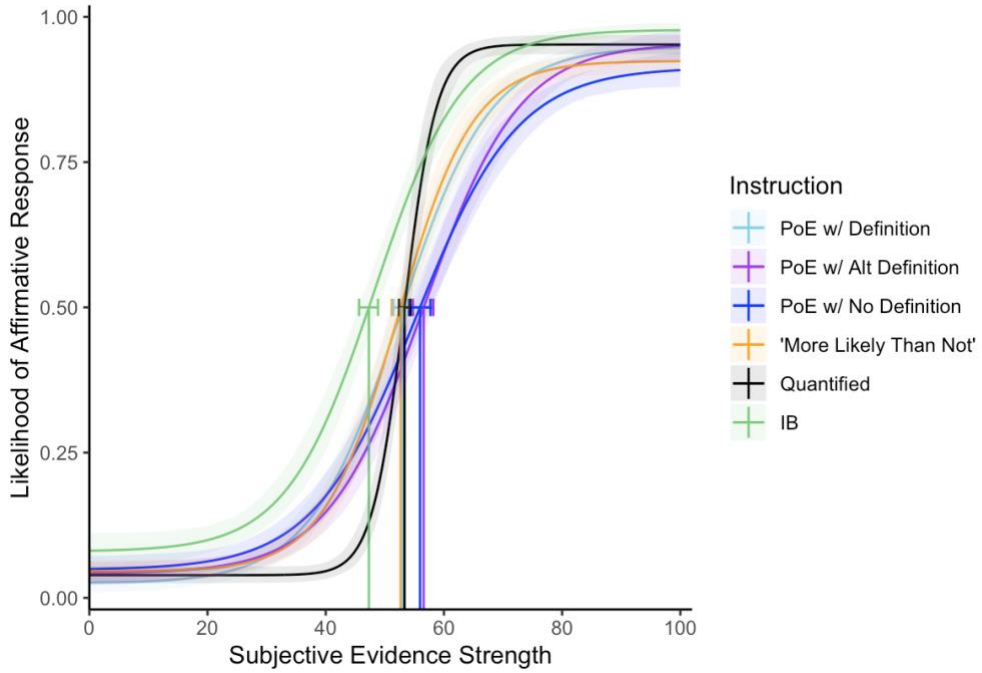

Figure 2. Likelihood of an affirmative response by subjective evidence strength and instruction. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Another interesting finding for instruction type is that the slope parameter for the quantified instruction was significantly steeper than all other instructions (vs PoE w/ Definition: -0.24[-0.45, -0.12]; vs PoE w/ Alt Definition: -0.24[-0.46, -0.14]; vs PoE w/ No Definition: -0.25[-0.46, -0.13]; vs "More Likely True Than Not": -0.21[-0.42, -0.10]; vs IB: -0.23[-0.44, -0.13]). This indicates that the instruction resulted in a sharp tipping point between yes and no responses, which makes sense as participants could compare the numeric value of the evidence strength to the quantified instruction (>50%) to reach a decision.

We then assessed the relationship between instruction and domain by comparing domain within each instruction type (Fig. 3) as well as comparing instruction within each domain (Fig. 4). Figure 5 presents the psychometric parameters for each function (instruction x domain; see Tables S10-S18 for all parameters and comparisons).

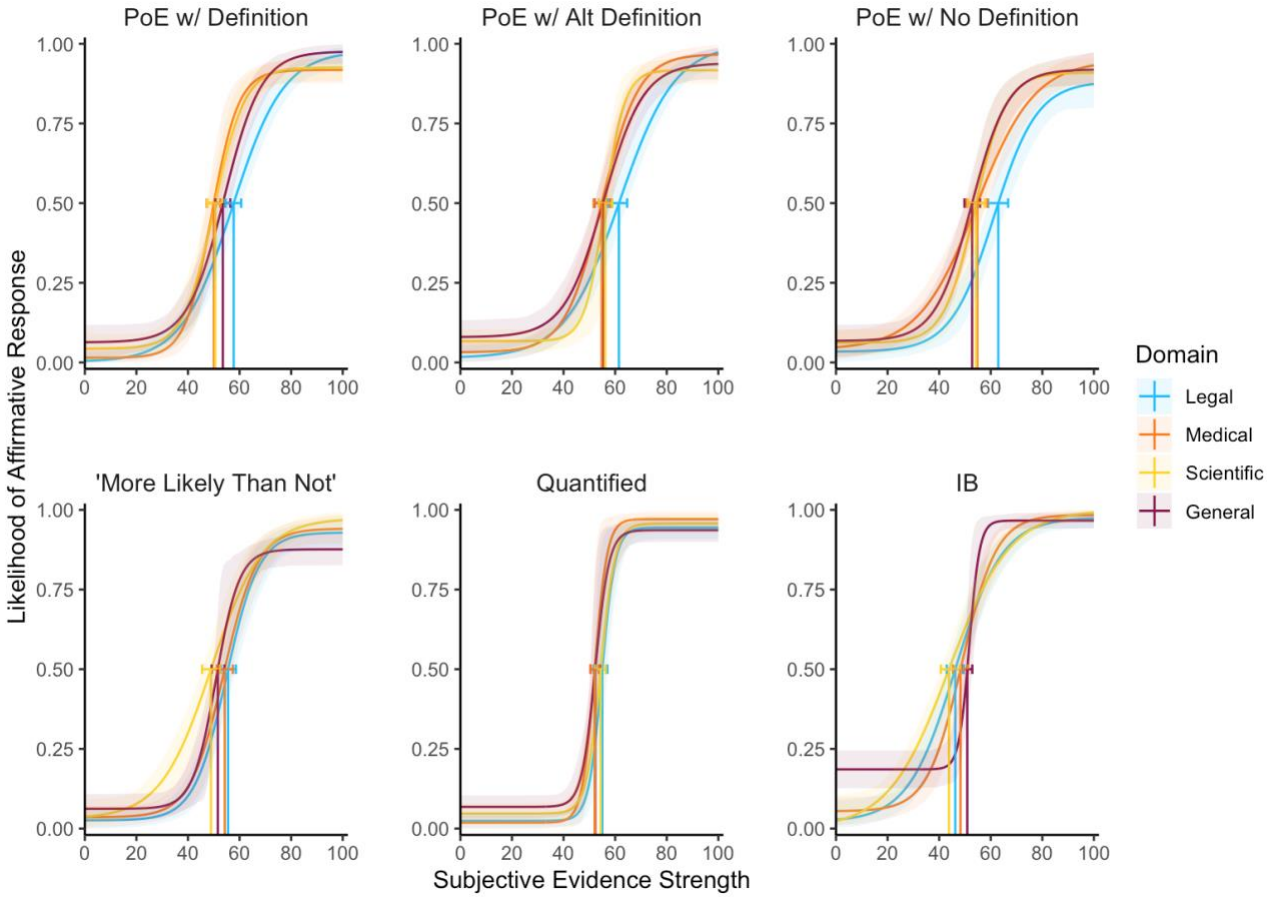

Figure 3. Likelihood of an affirmative response by subjective evidence strength and domain within each instruction. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

The effect of domain (i.e. legal > non-legal) was driven primarily by more stringent decisions in the legal domain versus the non-legal domains within the PoE instructions (cyan legal function in top row of Fig. 3, Fig. 5A). Specifically, decision thresholds were significantly greater in the legal domain compared to the medical and scientific domain for PoE w/ Definition

(Legal vs Medical: 7.68[1.36, 12.84]; Legal vs Scientific: 7.16[0.85, 12.74]), and significantly greater in the legal domain than all non-legal domains for PoE w/ No Definition (Legal vs Medical: 8.14[0.82, 15.73]; Legal vs Scientific: 8.64[1.16, 14.82]; Legal vs General: 10.15[3.50, 16.27]). The decision threshold for the "More Likely True Than Not" instruction was also significantly greater in the legal domain versus the scientific domain (6.64[0.64, 12.61]). Interestingly, we did not observe more stringent decisions in the legal domain for IB as we did in Chapter 2 (Experiment 1); indeed, decisions were quite liberal for this condition with a decision threshold lower than 50% (46.25[42.93, 49.50]). Ultimately, stringency within the legal domain relative to non-legal domains was observed mainly when paired with an instruction using the term PoE, which supports the idea that the overly stringent application of this standard is context dependent.

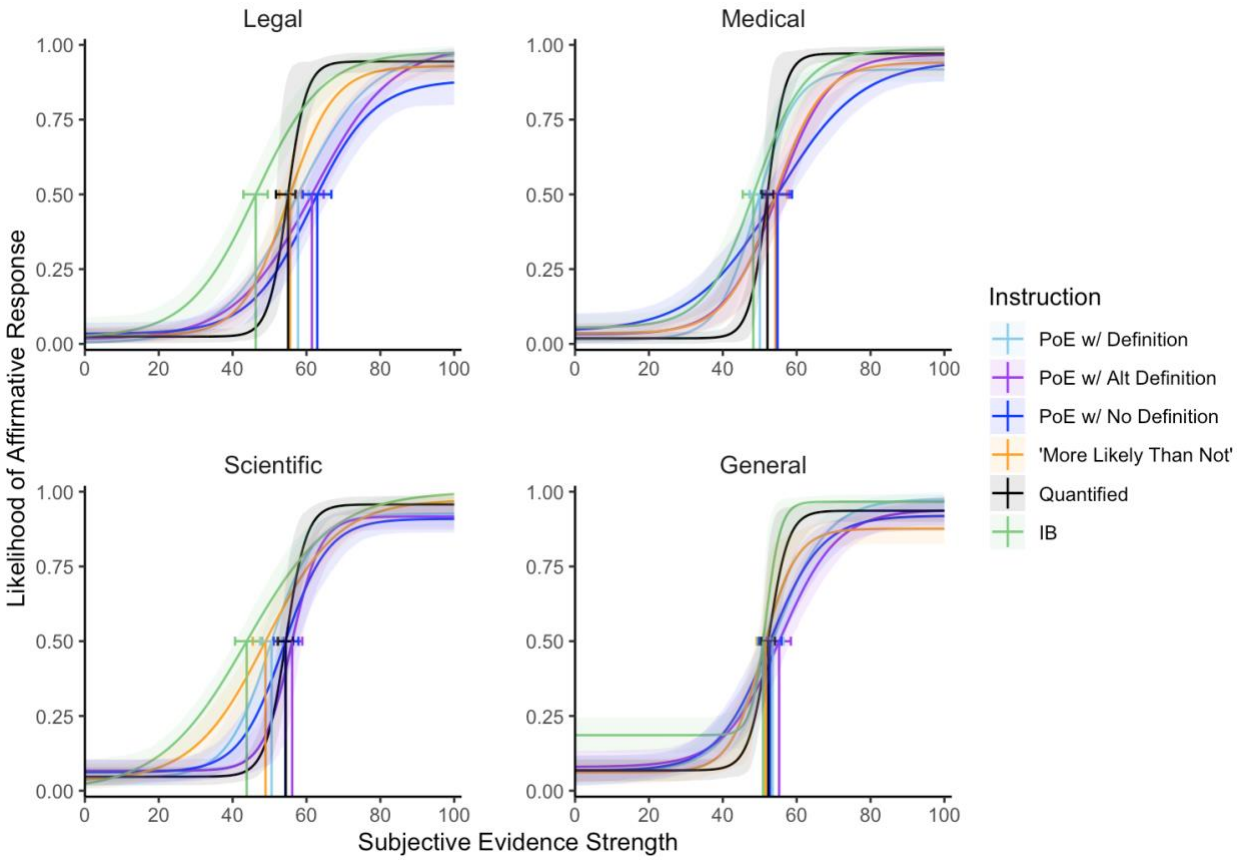

Figure 4. Likelihood of an affirmative response by subjective evidence strength and instruction within each domain. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Comparing instruction type within each domain allowed us to determine whether our five instructed conditions lead to similar decisions, as they all are equivalent to a theoretical decision threshold of just over 50% (Fig. 4, Fig. 5A). We found that for all of the non-legal domains, there was no significant difference for decision thresholds between instructed conditions with the exception of the PoE w/ Alt Definition versus "More Likely True Than Not" within the scientific domain (7.22[0.79, 13.20]). Within the legal domain, the PoE w/ Alt Definition and PoE w/ No Definition instructions were both significantly greater than the quantified threshold (PoE w/ Alt Definition vs Quantified: 6.44[0.29, 12.20]; PoE w/ No Definition vs Quantified: 7.91[0.93, 15.44]). This again provides partial support for our hypothesis that the overly stringent application of PoE is largely context dependent.

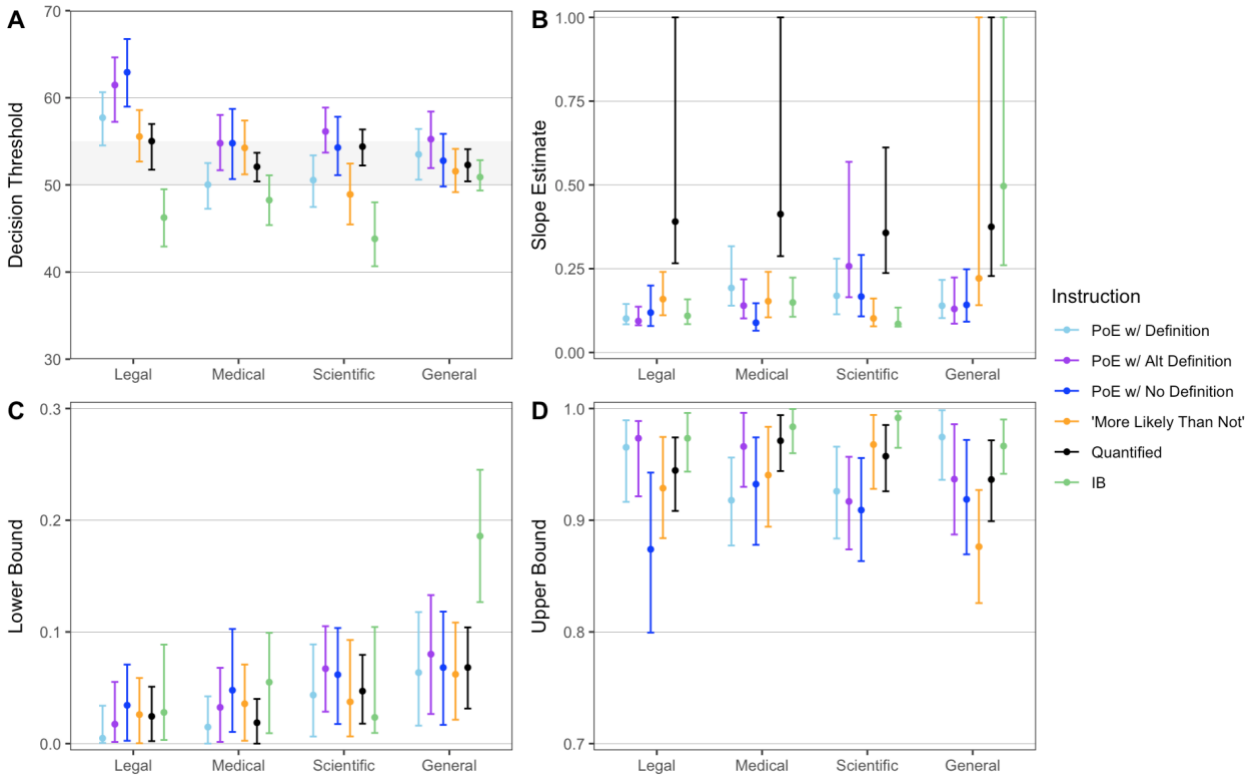

Figure 5. Subjective evidence strength at the decision thresholds (A), slope estimates (B), lower bound (C) and upper bound (D) by instruction type (colors) and domain (x-axes). 95% confidence intervals estimated via 1000 bootstrap samples. The large error bars for some of the slope conditions in B are due to a steeper slope that lead to some bootstrap samples having more of a step-function like shape, resulting in the increased upper part of the CIs.

Finally, we were interested in comparing instruction decision thresholds not just to each other but also to the prescribed standard for PoE; while theoretically this represents a tipping point of just over 50%, legal experts have been found to place PoE between 50% and 55% (McCauliff, 1982; Simon & Mahan, 1971, see also Chapter 3 of the present thesis). We therefore assessed the 95% confidence interval for the decision threshold of each condition (instruction x domain) to determine whether it included a value between 50 and 55 (Fig. 5A- gray shaded bar shows 50-55%; Table S10). The only instructed decision thresholds that were significantly greater than this range were the PoE w/ Alt Definition (61.46 95% CI[57.23, 64.64]) and PoE w/ No Definition (62.93 95% CI[58.98, 66.75]) within the legal domain. This suggests that the overly stringent

application of the PoE standard is specific to the legal domain and interacts with the language of the PoE instruction.

Discussion

In Chapter 4 we applied our psychometric approach to compare multiple theoretical equivalents of the PoE standard across societal domains to evaluate the factors underlying the overly stringent application of PoE instructions within the legal domain by laypeople (as seen in Chapters 2 and 3). The primary take away of this experiment is that it is the interaction of both a legal context and the language of a PoE instruction that results in laypeople applying this standard too stringently.

As noted in the Introduction, there have been two main interpretations of laypeople's overly stringent interpretation of the PoE standard: 1) individuals adjust the PoE standard upward (and BaRD downward) so that they align more closely with their intuitive decision threshold (e.g. Simon & Mahan, 1971); 2) non-experts are confused about what the legal burdens of proof mean, leading to increased variance in its application (Kagehiro & Stanton, 1985). Our results do not support either conclusion; across the results of Chapters 2-4 for non-experts, we have found that participants' decision thresholds are significantly more lenient for IB than PoE within the legal domain, indicating that people are not interpreting the PoE standard to better match their own intuitive decision threshold. Furthermore, within the non-legal domains of Chapters 2-4, we consistently found that laypeople are able to apply the PoE instruction in a manner consistent with the prescribed threshold, which suggests that at least in certain contexts individuals do properly understand PoE.

In addition to replicating the consistent findings described above, Chapter 4 critically compared five instruction types that should all be equivalent: three PoE instructions (PoE w/ Definition from Chapters 2-3, PoE w/ Alternate Definition, and PoE w/ No Definition), as well as "More Likely True Than Not" and a quantified instruction (>50%). For the non-legal domains, we found that decision thresholds for these instructed conditions were not significantly different from one another both within each domain (except scientific PoE w/ Alt Definition vs "More Likely True Than Not" as noted) and also for each instruction between domains, and furthermore that all thresholds were consistent with the prescribed standard. This is perhaps unsurprising given the consistent application of the PoE standard for non-legal domains. In contrast, several interesting findings emerged for the legal domain. First, the decision thresholds for "More Likely True Than Not" and quantified (i.e. the two instructions without the phrase "PoE") were consistent with both the prescribed standard and with the decision thresholds for the same instructions within the non-legal domains (with the exception of "More Likely True Than Not" for legal vs scientific). This indicates that participants are able to apply a threshold of approximately 50-55% in a legal context, and are not more stringent for PoE because they reject the threshold associated with the standard. Second, the decision thresholds for the three PoE instructions all demonstrated increased stringency (in relation to other instructions within the legal domain and/or compared to PoE instructions across domains) and were qualitatively the three highest thresholds. While the threshold for PoE w/ our original definition was significantly more stringent for the legal domain than either the medical or scientific domains, we note that it was not significantly different from the 50-55% standard (57.71 95% CI[54.53, 60.64]). The thresholds for PoE w/ Alternate Definition (61.46 95% CI[57.23,64.64]) and PoE w/ No Definition (62.93 95% CI[58.98, 66.75) were both significantly greater than this standard, and the PoE w/ No Definition was significantly greater for

the legal domain compared to all non-legal domains. This suggests that it is the interaction of the legal domain with an instruction containing the language "PoE" that leads to the overly stringent application of the PoE standard; indeed, only when both elements are present does this overly stringent application occur for PoE.

Given that the effect of this interaction is most pronounced for the PoE w/ No Definition instruction (Fig. 3), it is possible that participants are confused about what the PoE instruction means in all domains, but tend to rely more on their own intuitive decision threshold of around 50% and/or are more likely to make a 50-50 guess within the non-legal domains. However, given that instruction was manipulated between-subjects for this experiment, it is interesting that the same participants showed such a difference in the application of the same instruction. What is clear is that aspects of the legal domain lead individuals to be more conservative in their decision-making, even when presented with identical instructions and evidence strengths. It may be that individuals infer negative consequences for the defendant within the legal domain (i.e. punishment- fines or incarceration), and this causes them to apply more stringent decision thresholds. Correspondingly, past research has found that the potential punishment facing the defendant may influence verdict decisions, but these findings have been confounded with the severity of the crime/charge, the harm to victim, and the evidence associated with the crime (Freedman et al., 1994; Kerr, 1978; McComas & Noll, 1974; Vidmar, 1972). Chapter 5 uses a psychometric approach to examine the effect of the potential consequence/cost associated with a decision, with the goal of understanding the increased stringency of decisions within the legal domain observed in Chapters 2-4. Whatever may be the case, the present findings illustrate the critical importance of instructions and context in decision-making, and further suggest that the justice system should consider the adoption of definitions that do not contain the less familiar

legalese “preponderance of the evidence” such as “more likely true than not” to realize the prescribed decision threshold. At the very least, the present study calls for replication of the main findings in more realistic, mock cases.

CHAPTER 5

EFFECT OF DECISION OUTCOME COSTS

Introduction

In Chapters 2-4 we consistently found that individuals made more stringent decisions in the legal domain versus non-legal domains. Here we explore the possibility that this is due to participants applying a higher decision threshold in the legal domain because of the implicit high-stakes consequences of making such decisions, namely the punitive financial and/or incarceration outcomes to the defendant. Specifically, in Chapter 5 we assess the impact of decision costs in legal and, for comparison, in non-legal domains by explicitly manipulating the cost associated with participants' decisions.

Culpability and sentencing decisions are often dissociated in the US justice system. While civil juries are typically asked to make both liability and punishment/compensation decisions, the role of jurors in criminal cases is limited to reaching a verdict, and they do not determine the amount of punishment that a defendant receives (with exceptions in some US states and death penalty cases; Hans et al., 2015). It is the judges who typically make sentencing decisions based on established sentencing guidelines and mandatory minimums. Indeed, most states impose a general rule that the jury is not to be informed of the potential sentencing options facing an offender, and jurors may even be instructed to not consider possible sentences during their deliberations (Shannon v. US, 1994; US v. Chesney, 1996; both as cited in Barkow, 2003). However, it is common knowledge that more severe crimes generally call for harsher punishment, and it is reasonable to expect that individuals are aware that defendants will face consequences if

found guilty. Consistent with this notion, there is evidence that conviction rates are lower for more severe crimes/charges (McComas & Noll, 1974; Vidmar, 1972), as well as for more severe punishment (Kerr, 1978). These three studies were limited, however, by their use of a single case trial that presented strong evidence for manslaughter but then asked subjects to make verdict decisions for the more severe charges of first and second degree murder. Given the charges were disproportionate to the evidence, it is not surprising that the participants would be less inclined to punish. Indeed, a subsequent study that controlled for this confound did not find an effect of charge severity or punishment on conviction rates (Freedman et al., 1994). These studies aside, empirical research on the influence of punishment information on legal decision-making is quite limited, perhaps because of the challenges in manipulating charge severity and punishment in a realistic manner (i.e. presenting punishment that could reasonably be associated with the charge). Our psychometric approach may be particularly advantageous for exploring the relationship between decision outcome and punishment costs as it lends itself to a parametric manipulation of key variables and quantitative evaluation of their specific effects, as exemplified in Chapters 2-4.

In Chapter 5 we explicitly informed participants of what the outcome of their decision will be before they render it, and manipulated the cost associated with the decision outcome both within and between our scenarios. We also assessed how the effect of cost is influenced by the contextual domain and by who the cost affects. Specifically, given our interest in the cost of legal decisions (which typically affect an individual) and to assess whether the effect of such cost on decisions is unique to the legal domain, we compared these to costs to an individual in the medical domain, to costs to an individual in miscellaneous domains (i.e. general domain), and also to costs to a community in the general domain. We selected the medical domain as control because it is one in which medical decisions routinely have important consequences to the individual. And to have a

broader perspective on the effect of decision outcomes on decisions, we also included conditions that manipulated costs to an individual or to a group of individuals in miscellaneous (general) domains.

Methods

Participants

We recruited participants from the United States via Amazon Mechanical Turk (44% male, 56% female, 24 non-binary; Mean age=38.00 years, range=18-89). As in the previous chapters, we recruited participants until we reached roughly 40 observations per cell after all exclusions. Participants who successfully completed the survey were paid $0.75. The average time to completion was between five and eight minutes on average. All participants provided informed consent, and the experimental protocol was approved by the Vanderbilt University Institutional Review Board.

Design and Materials

The task employed a 4 (domain: legal, medical, scientific, control; within-subjects) x 9 (objective evidence strength: 0%, 20%, 40%, 60%, 80%, 90%, 95%, 99%, 100%; within subjects) x 3 (decision criteria instruction: IB, PoE, BaRD; within-subjects) x 2 (decision cost: low, high; within-subjects) design. Participants responded to four scenarios, one scenario from each domain in random order, with random assignment to scenario, objective evidence strength, decision criteria instruction, and cost level. After these four trial scenarios, participants completed an attention

check scenario that was identical to the trial scenarios in its design but contained specific language instructing participants to provide a specific response on the next screen (see Supplementary).

Participants completed the study on the Qualtrics online survey platform at their own pace. The initial presentation of the scenario and instruction for each trial was identical to the previous chapters; participants read the scenario and clicked a button to continue, at which point the decision criterion instructions (for the PoE and BaRD conditions) appeared below the scenario on the same page. The language for the instruction types was identical to Chapters 2 and 3 (see Chapter 2 Fig. 2). Participants were randomly assigned to the IB, PoE, or BaRD instruction for each scenario. For this experiment, however, after participants read the instructions and clicked to continue (or right after the scenario for those in the IB condition), two outcome-related sentences appeared on the screen followed by the decision prompt (see Fig. 1). The first sentence stated the outcome if the participant gave an affirmative response (e.g. "If you believe that Mark stole the company's data, he will be required to pay a $10,000 fine"); this sentence always included the cost to the person(s) in the scenario, which was manipulated as low or high (see below). The second sentence stated the outcome if the participant did not give an affirmative response, which was always that nothing would occur (e.g. "If you do not believe that Mark stole the company data, no action will be taken against Mark"). These two sentences were followed by the decision prompt (e.g. "Do you believe that Mark stole the company's data?", "Do you believe that Mark will develop Huntington's disease?"). Participants read the potential decision outcomes and decision prompt and made a yes/no response. On a new page they then provided their own subjective probability for the event occurring (e.g. "What do you believe is the probability that Mark stole the company data?") as in the previous experiments.

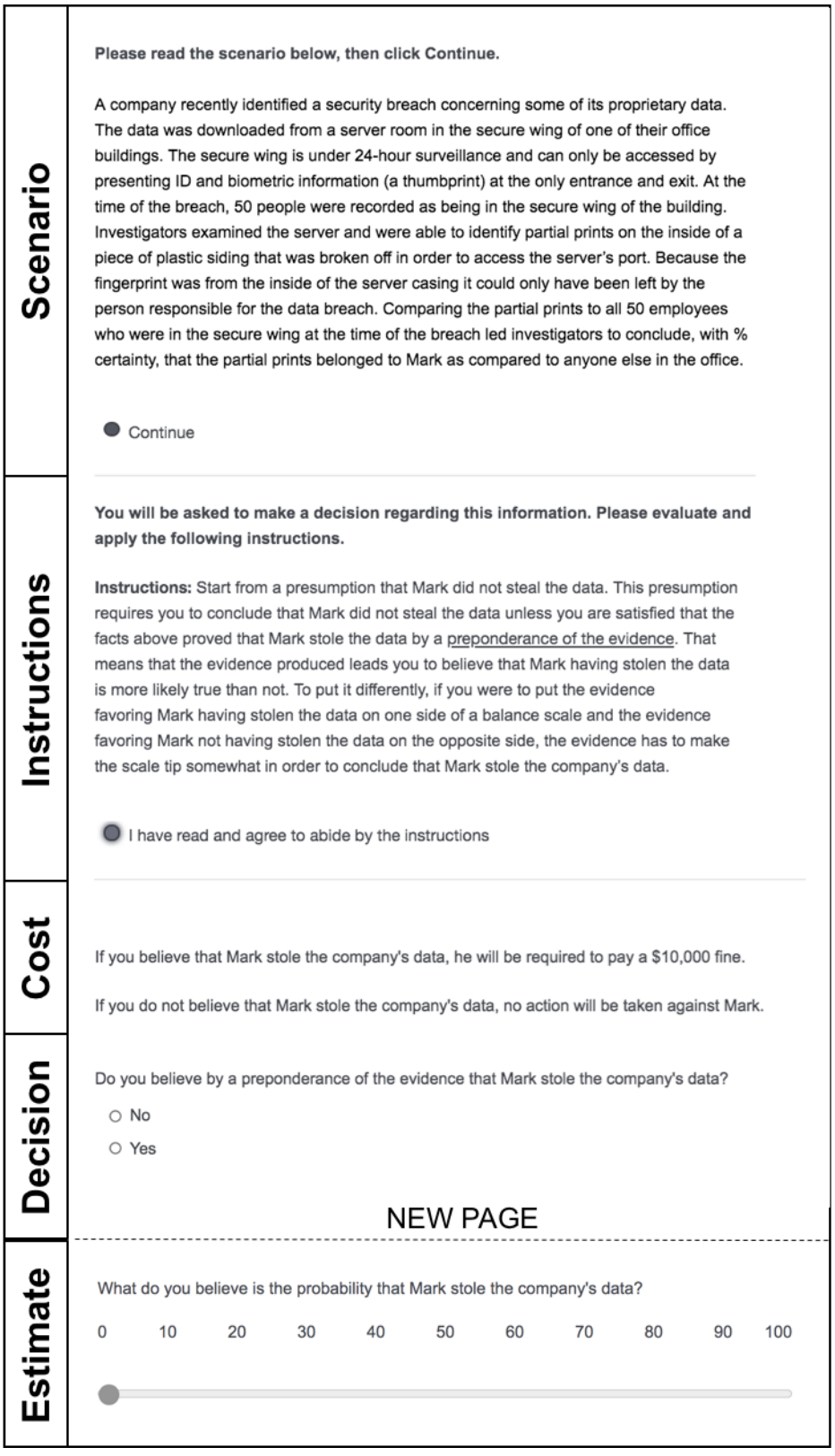

Figure 1. Sample trial as seen by participants in the legal domain (company data theft x fingerprint evidence scenario) with PoE instruction and low level cost.

Given our hypothesis that decisions may be more stringent in the legal domain because participants infer a punishment cost for the scenario's protagonist, we adopted the scenarios in the non-legal domains for the present experiment to allow us to more directly compare them with the legal scenarios. First, some of the costs in the non-legal domains were specific to an individual, as in the legal domain, whereas some others were to a community of individuals. Specifically, the experiment included the following domains: Individual Legal- associated with a punitive cost affecting an individual; Individual Medical- associated with the cost of medical treatment affecting an individual; Individual General- associated with costs affecting an individual outside of a legal or medical context; and Community General- associated with costs affecting a community outside of a legal or medical context (adapted from the scientific domain scenarios).

The legal and non-legal scenarios were also modified to include comparable ranges in cost outcomes across domains. Each scenario was associated with an outcome for an affirmative response and an outcome for a negative response (see Fig. 1). The outcome for a negative response was always that no action would occur. The cost of the outcome for an affirmative response had two levels, low or high, which were selected so as to be realistic with regards to the scenario content, with the high cost level always roughly 10x the cost of the low cost level (e.g. fine of $10,000 vs $100,000). (We note that the designation of a cost as low versus high was only in relation to the levels within each scenario- the low level cost of one scenario is not necessarily low in relation to the costs of other scenarios, as the costs within a scenario had to be realistic and we intentionally presented a wide range of cost values across scenarios). We included both monetary costs and costs that represented a loss to the individual in temporal terms (e.g. duration of incarceration, length of physical incapacitation following treatment), with at least one monetary and one duration cost in each domain. The monetary costs ranged from $5,000, $10,000, $50,000,

$100,000, to $1,000,000. Importantly, some of these monetary costs (i.e. $10,000 and $100,000 costs) were common across domains, thus allowing for a direct comparison between domains for the same cost. The duration costs ranged from 1 Day, 10 Days, 1 Month, 3 Months, 6 Months, 1 Year, 1.5 Years, 5 Years, 10 Years, to 100 Years. For the temporal costs, we had common costs between the individual general and community general domains (i.e. 1 Day and 10 Days) but no others due to the challenge in creating a realistic range of time within each scenario. All scenarios and their possible outcomes are included in the Supplementary Materials.

Individual Legal scenarios were identical to those of the legal domain used in Chapters 2-4. We had three fact patterns (stealing company data, stealing prescription drugs, and murder) crossed with three types of evidence (video facial recognition, finger prints, and DNA), with the objective evidence strength as the level of certainty with which investigators concluded that the evidence was left by the protagonist, Mark, presented as a frequentist measure of probability with random assignment to one of nine levels within-subject: 0%, 20%, 40%, 60%, 80%, 90%, 95%, 99%, 100%. The cost levels for each fact pattern were as follows: stealing company data- $10,000 versus $100,000 fine; stealing prescription drugs- 6 months versus 5 years in prison; murder- 10 versus 100 years in prison.

Individual Medical scenarios were the same three fact patterns as those of the medical domain used in Chapters 3-4 but with a named protagonist, Mark, rather than referring to a "patient" to make it clear that the cost was to a specific individual, as in the Individual Legal scenarios. The objective evidence strength was again presented as a frequentist probability (same levels as above) for the level of certainty of Mark having the disease given the test result. The cost levels for each fact pattern were as follows: Huntington's disease- $10,000 versus $100,000 per

year out-of-pocket treatment; Irritable Bowel Disease (IBD)- 1 month versus 1 year liquid diet following treatment; optic nerve damage- $5,000 versus $50,000 per year out-of-pocket treatment.

Individual General scenarios consisted of three fact patterns describing the likelihood (i.e. objective evidence strength as frequentist probability) of either: Mark's small business needing to purchase specific liability insurance within the next year based on a risk management assessment, an incoming packet needed for Mark's work containing a virus based on detection software, or Mark's house remaining on the market for 6 months based on features of the house. The cost levels for each fact pattern were as follows: Insurance: $10,000 versus $100,000 per year additional liability insurance; Virus: 1 day versus 10 days unable to access the packet for work; Real Estate: $10,000 versus $100,000 decrease in the asking price of the house.

Community General scenarios consisted of three fact patterns describing the likelihood (i.e. objective evidence strength as frequentist probability) of either: above average temperatures developing in the Pacific Ocean (which would threaten local fish populations) based on meteorological patterns, a lake being unsafe for swimming based on the sulfate content, or interstellar debris damaging a telecommunications satellite based on astronomical measurements. The cost levels for each fact pattern were as follows: Water Temperature- 3 months versus 1.5 years restricted access and fishing for local coastal region; Lake Water- $100,000 versus $1,000,000 cost to lake residents to resorb sulfates; Interstellar Debris- 1 Day versus 10 Days widespread US mobile phone disruption due to moving the satellite out of the path of the debris.

Statistical Analyses

We used the same procedures described in detail in Chapter 2 to fit psychometric curves, estimate parameters, obtain 95% confidence intervals, and perform comparisons between

conditions using 1000 bootstrap samples. The confidence intervals for all pairwise comparisons were Bonferroni-corrected, and as with the previous experiments a confidence interval that does not contain 0 indicates a significant difference.

Results

We first compared decisions between domains (Fig. 2, see Table S1 for all parameter comparisons), collapsing across decision criterion instructions and decision costs. Consistent with the previous chapters, decisions were most stringent in the legal domain (Fig. 2), with a decision threshold significantly greater than all other domains (vs Medical: 9.54[7.04, 11.85]; vs General: 13.28[10.87, 15.64]; vs Community: 15.70[12.91, 18.18]. Interestingly, decisions also differed between non-legal domains such that after the legal domain, decisions were most stringent for the medical domain followed by individual general and then community general. This was borne out in statistically significant differences between each of these domains (Medical vs Individual General: 3.74[1.43, 6.31]; Medical vs Community General: 6.16[3.58, 8.71]; Individual General vs Community General: 2.42[0.004, 4.80]).

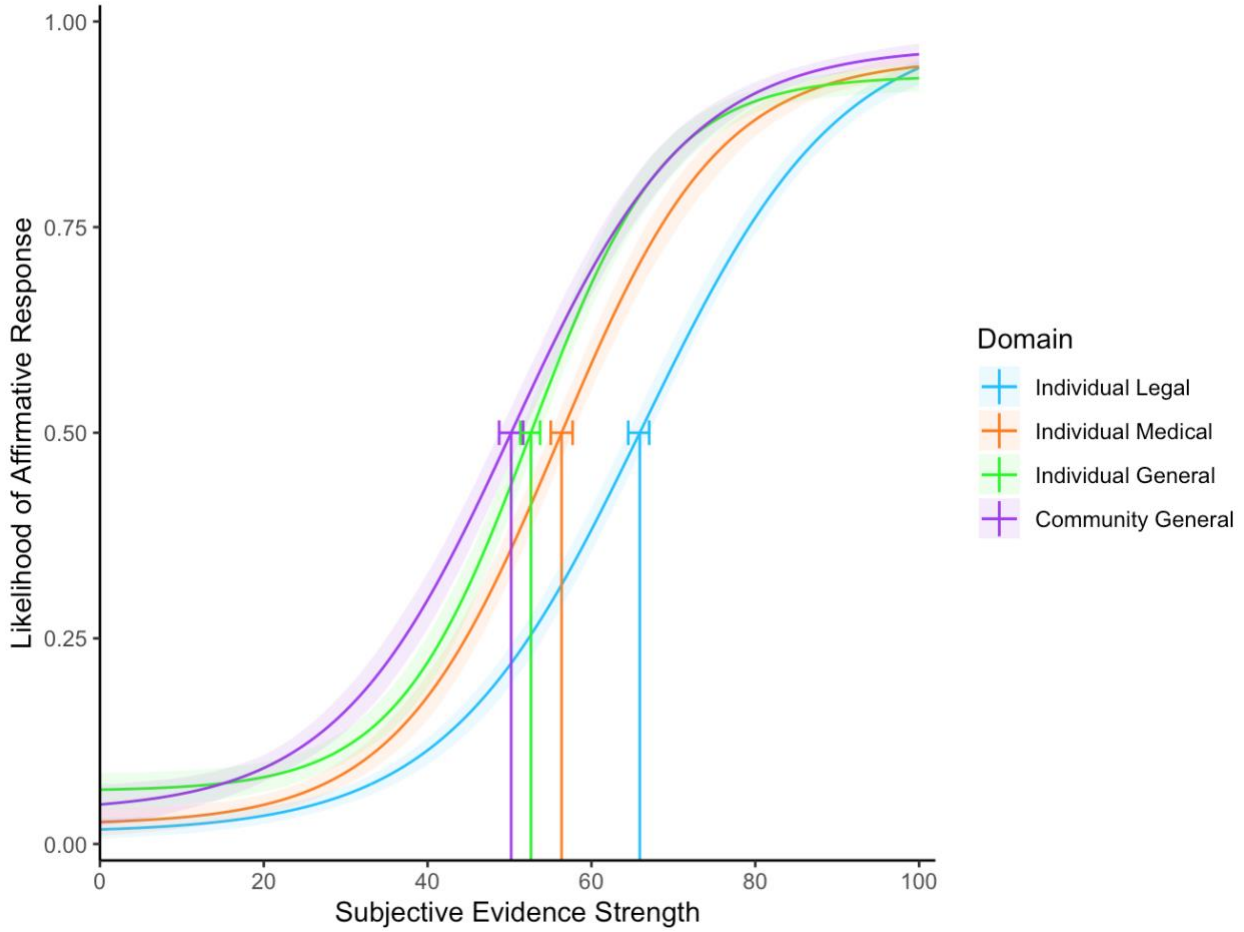

Figure 2. Likelihood of an affirmative response by subjective evidence strength and domain. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Next we assessed the relationship between decision criterion instruction and domain by comparing domains within each instruction type (Fig. 3) as well as by comparing instructions within each domain (Fig. 4; Table S2). The difference and Bonferroni-corrected confidence intervals for the significant comparisons discussed below are presented in Table S2; we do not list them in the text due to the high number of significant comparisons.

Comparable to the data collapsed across instruction type (Fig. 2), decision thresholds for each instruction type were significantly greater in the legal domain versus all others, and the medical thresholds were significantly greater than the individual general thresholds (Fig. 3, Table

S2). The individual general threshold was significantly greater than the community general threshold for the PoE instruction. The decision threshold for the legal domain with IB instructions is noteworthy (Fig. 4), as it was more stringent than in our previous experiments which all used the same scenarios. This suggests that the addition of explicit decision costs lead to more conservative decisions in the legal domain even in the absence of a legal instruction. Decision thresholds were significantly different between IB, PoE and BaRD for all domains such that IB<PoE<BaRD (Fig. 4, Table S2). These differences were not significant in the non-legal domains for IB versus PoE for our non-experts in Chapter 3, though it appears that for the individual general and community general domains this is due to lenient decisions for the IB instruction; this suggests that participants favored a false positive for those domains in the absence of any instruction.

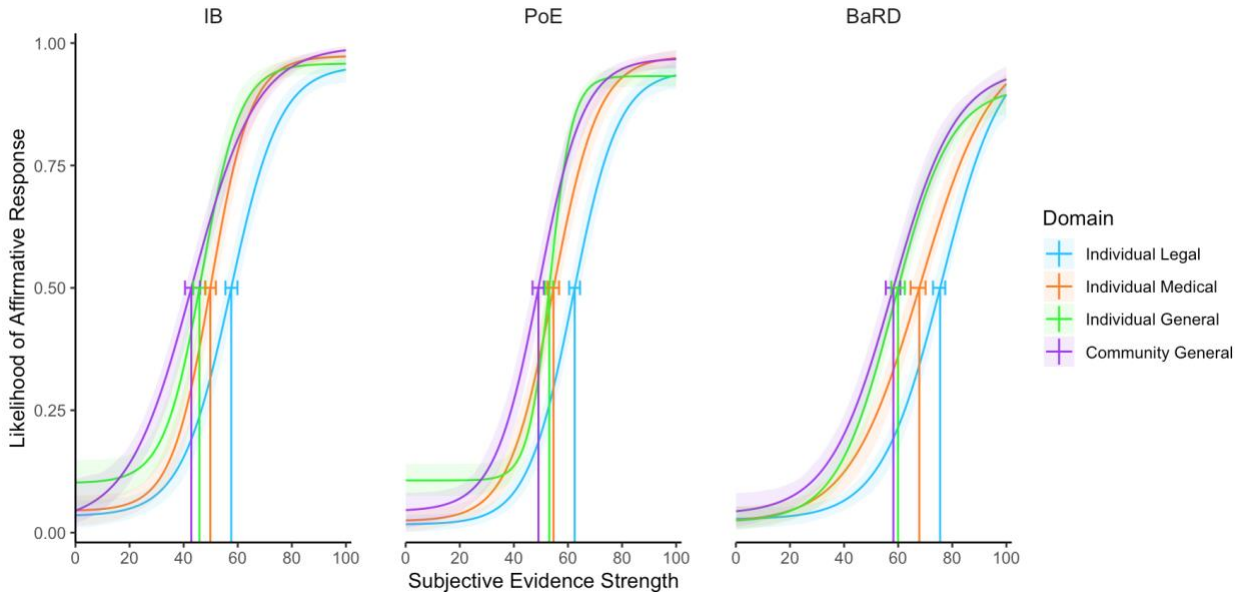

Figure 3. Likelihood of an affirmative response by subjective evidence strength and domain within instruction. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

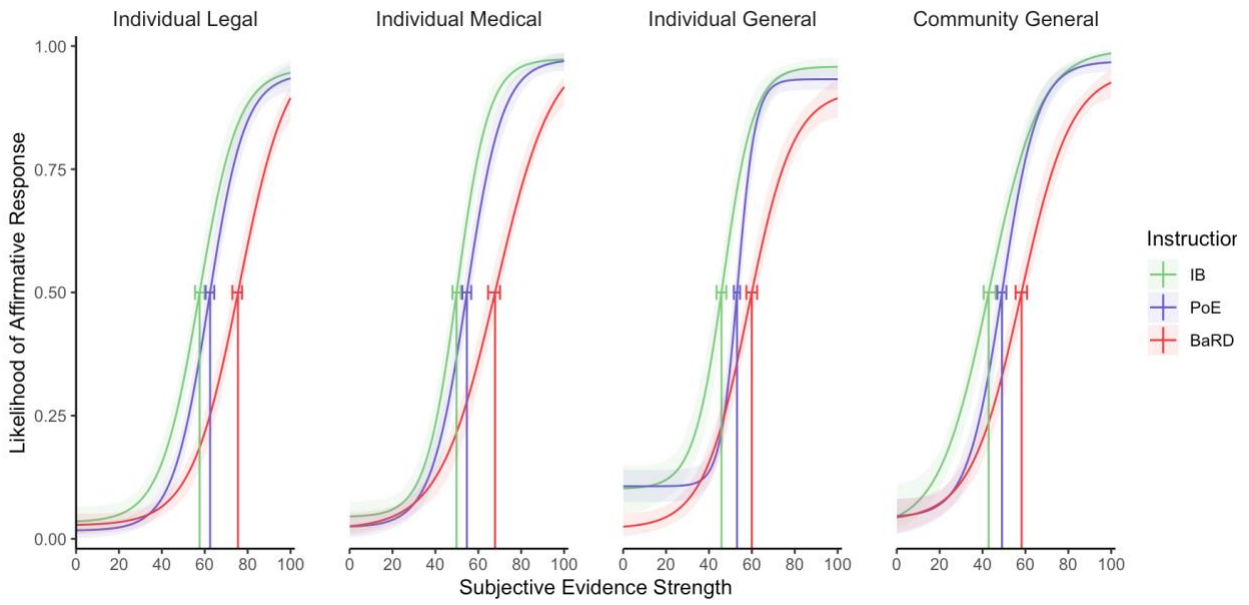

Figure 4. Likelihood of an affirmative response by subjective evidence strength and instruction within domain. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

Surprisingly, we did not observe an effect of low versus high cost within domain (Fig. 5). The only differences were a steeper slope for high versus low cost in the legal domain (-0.04[-0.07, -0.002]) and a lower upper bound for high versus low cost, also in the legal domain (0.09[0.04, 0.14]). As noted in the Methods, the designation of a cost as low versus high for each scenario was relative to each scenario but the low cost in one scenario could be greater than the high cost of another scenario within the same domain (e.g. low level cost for murder=10 years in prison, high level cost for prescription drug theft=5 years in prison). To further assess the effect of cost within each domain we therefore used the absolute cost amount provided to participants; each domain had six cost amounts (three fact patterns w/ two cost levels each). Cost amount had little impact on decisions (see Figs S3-S6 and Tables S3-S6). This may not be surprising, however, since this compares different kinds of costs (monetary and duration). Surprisingly, however, this null finding held even between low and high cost levels within single scenarios (see Figs S3-S6 and Tables S3-S6). Overall, decision thresholds were not significantly different between absolute

cost amounts within the legal, medical, or general domains (with the exception that within the individual general domain there was a greater decision threshold for the $100,000 Real Estate versus the $10,000 Liability Insurance scenarios (-7.20[-13.38, -1.58]).

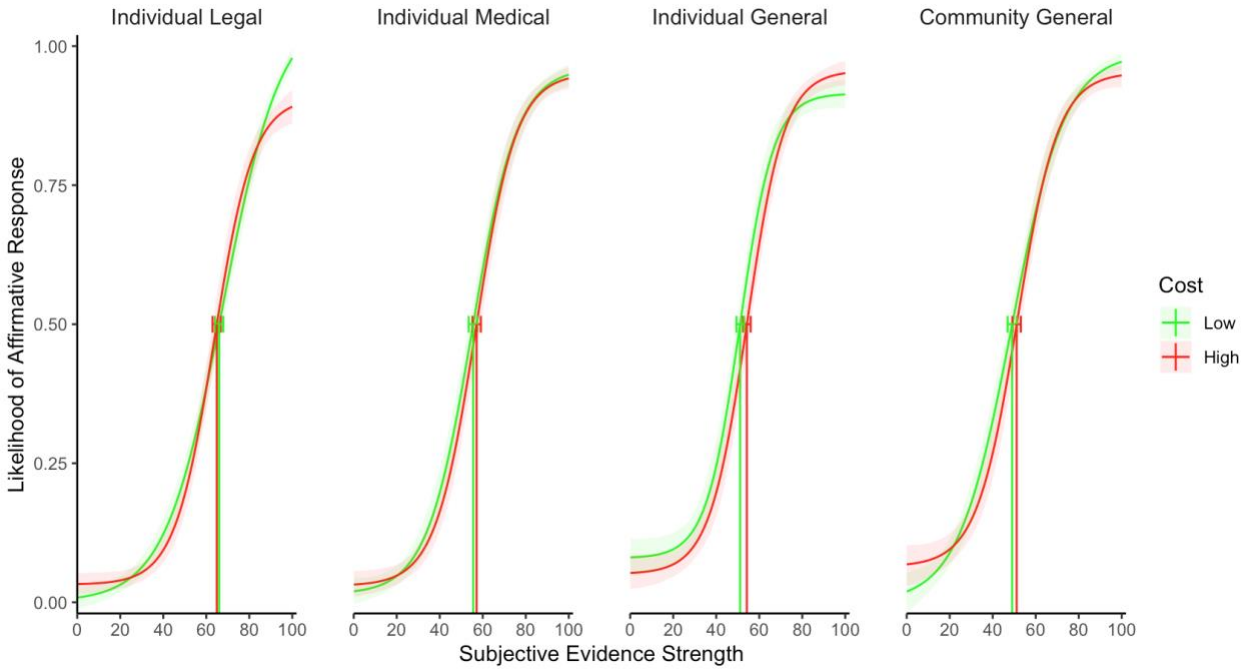

Figure 5. Likelihood of an affirmative response by subjective evidence strength and cost level within domain. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

To determine whether including any cost (low or high) influenced decisions, we compared decision parameters for the low and high costs in the legal domain to those in the legal domain from our non-expert controls in Chapter 3, who responded to identical legal scenarios but without any decision outcome cost given (the non-legal domain scenarios were modified/altered for the present study). Figure 6 shows the effect of cost by instruction within the legal domain. Interestingly, we did not observe differences between the low or high costs and the no cost condition except for between the high cost and no cost for the IB instruction. The high cost decision

threshold was greater (-5.68[-10.30, -0.88]) and the high cost upper bound was lower (0.09[0.04, 0.38]), indicating more conservative decisions for high costs compared to when no cost was given. It is striking however that there were no other differences for when participants received a cost compared to when they did not. This does not necessarily mean that cost doesn't matter at all, as participants may infer costs even when one is not explicitly provided.

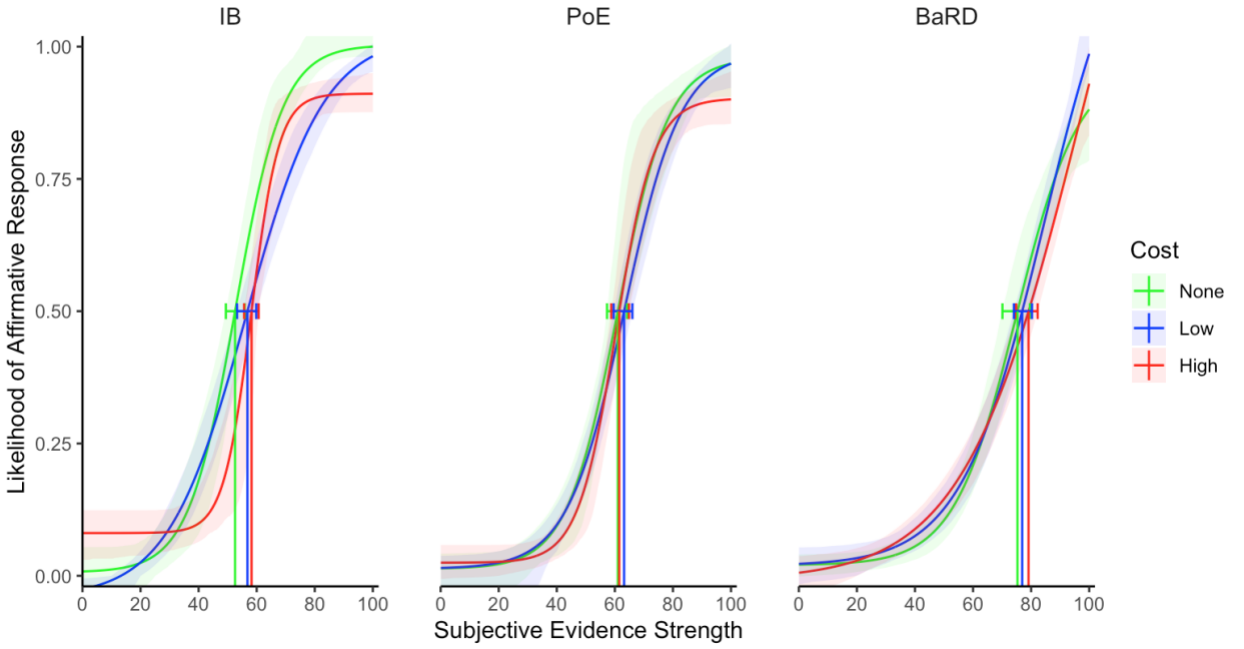

Figure 6. Likelihood of an affirmative response by subjective evidence strength and cost level within instruction. Shaded regions are 95% confidence intervals estimated via 1000 bootstrap samples. Decision thresholds are marked with vertical lines and 95% error bars.

To assess the effect of comparable decision costs onto decisions across domains, we next plotted all decision thresholds by ascending cost amount across domains in order to determine whether higher absolute costs in general are associated with differences in decisions (Fig. 5). First, this figure reveals how little the decision thresholds change with ascending monetary or time costs, consistent with the results described above. Instead, we observed the same trend as in Figure 2, with the most stringent decisions in the legal domain, followed by the medical, individual general,

and community general domains. Interestingly, the 6-month legal cost has a higher decision threshold than medical or general costs with longer times (Fig. 7). This may be explained by the fact that participants deem 6 months of incarceration as being far costlier than 1 year on a liquid diet. Comparison across domains is more appropriate when the monetary costs are comparable, however, which is the case for the $10,000 and $100,000 monetary costs (see Methods). Pairwise comparisons within the $10,000 cost amount found that the legal threshold was significantly greater than all of the non-legal threshold (Legal vs Medical: 11.25[5.04, 17.06]; Legal vs Individual General-Insurance: 18.41[12.06, 24.44]; Legal vs Individual General-Decreased Asking Price: 13.97[7.89, 19.61]), while the threshold for the medical domain was significantly greater than for individual general-insurance (7.16[0.71, 14.21]). Within the $100,000 cost amount, the legal threshold was significantly greater than the individual general-insurance (9.61[2.93, 16.91]), individual general- decreased asking price (6.60[1.36, 12.03]), and community general thresholds (14.52[7.56, 20.84]). The community general threshold was also significantly lower than the medical (9.32[1.72, 15.59]) and individual general-decreased asking price (7.92[1.79, 13.52]) thresholds. Thus, decision thresholds in the legal domain are higher than in other domains even when decision costs are explicitly equated across domains.

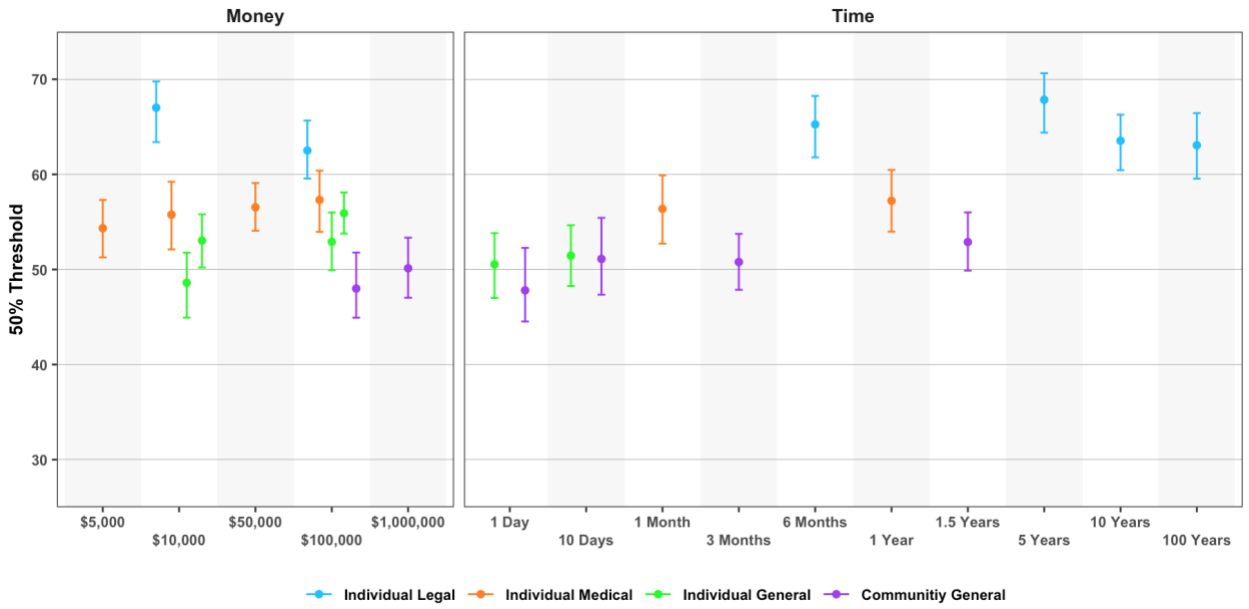

Figure 7. Decision thresholds with 95% confidence intervals by cost amount in ascending order. The left panel presents the monetary costs while the right presents time costs.

Discussion

In Chapter 5 we explicitly informed participants of the outcomes associated with their decisions and manipulated the cost within and between scenarios. We had hypothesized that participants' decisions are more stringent in the legal domain than in other domains because they infer that the defendant may face costly punishment. We compared costs to individuals in the legal domain with costs to individuals in the medical and general domain, as well as costs to communities in the general domain. The main upshot of this experiment is that laypeople are inherently more conservative in their decisions in the legal domain irrespective of the explicit decision cost amounts.

We expected that there would be an effect of cost amount such that decisions would be more stringent for higher costs, particularly within the legal domain. However, like Freedman et al. (1994) we did not observe an effect of punishment severity on decisions in the legal domain, nor an effect of cost amount on decisions within any non-legal domain despite the 10-fold

difference between low and high punishment costs. Even a wrongdoing associated with 6 months of incarceration led to the same decision threshold than a far more serious wrongdoing tied to a 100 years-long incarceration. The fact that participants were not exposed to the two levels of decision costs for the same scenario was likely important in the outcome of our results as it prevented reference effects in punishment costs.

It is striking that we found that the domain in which the decision is rendered is far more important in determining the decision threshold than is the cost of making that decision. This is most obvious in the comparison of monetary costs, because we presented the same cost amounts in multiple domains ($10,000 and $100,000) and found that even for the same cost to an individual, the legal domain still resulted in more stringent decision thresholds. Furthermore, this effect was so pronounced that a $10,000 punitive cost to an individual in the legal domain lead to more stringent decisions than $50,000, $100,000, and even $1,000,000 costs to individuals in other domains. Our findings suggest that participants may be more stringent in the legal domain not because they anticipate specific punishment consequences for the defendant, but rather because administering any sort of punitive cost leads to more conservative decisions. In other words, the legal domain may implicitly call for more stringent decision criteria simply because it is inherently associated with negative consequences (to the defendant) upon such decisions. In contrast, decision-making in other domains is not inherently or exclusively detrimental. For example, in the medical domain the cost of the treatment may lead to improved health outcomes for the patient. Indeed, the legal domain may be unique in that substantial costs (monetary, incarceration, even capital punishment) can be imposed on an individual without there being any potential benefit for the individual themselves (though some may argue the correctional value of incarceration). The legal scenarios in our experiments have revolved around the identity of the offender (i.e. there is

no question that the crime took place, only who did it); participants may be stringent in the legal domain because they are unwilling to impose a punitive cost on an innocent person, though it is interesting to note that we did not observe an effect of punishment costs even when participants used a very stringent decision criteria (BaRD).

We acknowledge several limitations to the present study. First, as in all the experiments throughout this thesis, the limited effect of punishment costs observed here is specific to the present, non-real word circumstances, and it would be interesting to see if these effects would be repeated in mock trials or real world situations. Future research could also explore different types of legal contexts to determine whether this stringency exists in the legal domain when there is no uncertainty of whether the defendant is guilty. Furthermore, presenting a wider range of potential costs could determine whether there is a point at which the decision cost has an effect, for instance by including more extreme differences between low and high level costs.

These caveats notwithstanding, we draw two important conclusions from the present study. First, that decision costs appear to have little impact on the decision criterion, not only in the legal domain, but in several other domains as well. Second, that legal decisions are rendered in a more conservative manner than decisions in other societal domains regardless of the actual punitive cost. In some ways, both may be welcome news to the legal system. Decisions about guilt should not, ideally, be influenced by the outcome of that decision, and legal decisions should be inherently conservative given the serious implications they have upon the life of the individual.

CHAPTER 6

EFFECT OF CHARACTER EVIDENCE AND ADMISSIBILITY INSTRUCTIONS ON TPP

Introduction

In Chapters 2-5 we examined the influence of legal burden of proof instructions on decisions using a psychometric approach. As stated in the General Introduction, in Chapter 6 we assessed another class of legal instructions that are commonly issued to jurors with the goal of reducing bias to render impartial and equitable decision- instructions to disregard evidence (Federal Rules of Evidence). When evidence that has been presented to a jury is deemed inadmissible judges may instruct jurors to disregard it, which means that jurors should not consider the evidence in order to reach a decision. As described below, findings related to the efficacy of disregard instructions are mixed, and it is unclear how jurors may attempt to suppress information in response to such instructions and how these instructions affect punishment decision processes. While a psychometric approach was advantageous in the previous chapters to quantify and compare decision parameters, in Chapter 6 we incorporate imaging data into understanding the impact of inadmissible character evidence and disregard instructions on punishment. Specifically, we use fMRI to examine how instructions to either consider or disregard character evidence during punishment decision-making affect the neural network underlying third-party punishment decisions. We build off of previous fMRI studies in the Marois lab that have identified this network and the brain mechanisms involved in third-party punishment.

There are a number of reasons that evidence may be ruled inadmissible; because it was obtained illegally, is hearsay, is irrelevant, or is prejudicial (Federal Rules of Evidence). While the

admissibility of evidence is typically determined in advance of the trial and away from a jury, it is not uncommon for inadmissible evidence to be presented to a jury when a lawyer or witness discloses evidence that has been ruled inadmissible or has not yet been ruled on. If the evidence is deemed inadmissible, the jury is instructed to disregard it, and this instruction is typically repeated to jurors just prior to deliberation. Of particular interest to the present study is character evidence, or evidence related to the defendant's character that shows a propensity to commit a crime, as it is considered inadmissible in criminal trials (Rule 404 of the Federal Rules of Evidence). Such evidence, including disclosing prior "crimes, wrongs, or other acts" is considered prejudicial as it may bias jurors against the defendant. Negative character traits have been found to predict higher guilt ratings and greater suggested sentences (Izzett & Leginski, 1974; Kaplan & Kemmerick, 1974; Landy & Aronson, 1969; Wissler & Saks, 1985), and jurors are more likely to convict when they hear evidence of prior convictions (Greene & Dodge, 1995).

While there is broad agreement that character evidence can be prejudicial and should not be admissible when rendering legal decisions, there is no such consensus about the efficacy of disregard instruction in achieving its intended goal. Some studies suggest that jurors can follow directions to disregard certain types of inadmissible evidence (Simon, 1966) or even overcompensate in correcting for the evidence (Thompson et al., 1981). Other literature, however, comes to the conclusion that jurors' efforts to ignore evidence is not equivalent to them having never heard it to begin with (Carretta & Moreland, 1983; Casper et al., 1989; Edwards & Bryan, 1997; Lieberman & Arndt, 2000; Sue, Smith, & Caldwell, 1973; Tanford & Cox, 1988). Individuals may be able to follow disregard under certain conditions, namely when they understand and agree with the reason for the inadmissible ruling, for certain types of evidence (e.g. hearsay), or when the evidence favors acquittal rather than conviction (Kassin & Sommers, 1997;

Thompson et al., 1981). Interestingly, a stronger admonishment of the jurors to disregard evidence or an attempt to provide legal reasoning can lead to higher rates of conviction and higher damages for the defendant (Broeder, 1959; Pickel, 1995; Thompson et al., 1981). It has been suggested that this back-firing of the disregard instruction is due to reactance by jurors in response to feeling that the judge is trying to tell them what to decide or reducing their decision freedom (Lieberman & Arndt, 2000; Thompson et al., 1981). However, there is stronger support for back-firing being due to the instruction inadvertently drawing attention to the inadmissible evidence and emphasizing to jurors that it is important in some way (Broeder, 1959; Lieberman &Arndt, 2000; Pickel, 1995).

While much work has focused on understanding the behavioral effects of introducing prejudicial evidence and of instructing to ignore it, far less is known about the brain mechanisms underlying these processes. Yet, elucidating the way in which character evidence – and the instructions to disregard it – impacts the brain circuitry supporting legal decision-making may shed light on the potential efficacy of disregard instructions. How is (biasing) character evidence encoded in the brain, how does it affect activity in the neural network involved in legal decision-making, and how does instruction to disregard this evidence modulate that activity? Specifically, do such instructions lead to an extinction of any neural trace of character evidence, or do instructions serve to actively suppress evidence-related activity? If it is the latter, it would suggest that biasing evidence may be harder to regulate and may surface at any time to influence legal decision-making. Answering these questions may lead to a better understanding of the effect of character evidence and instructions to disregard it, which in turn can ultimately lead to the implementation of informed practices or policies to effectively address this issue in the legal system.

The aim of the present study is to identify the imprints of character evidence in the human brain, characterize how it may modulate activity in the neural network underlying legal decision-making, and to understand how Disregard instructions affects that character-related activity. We address these aims by introducing character evidence and admissibility instructions into a previously used experimental paradigm to determine the influence of each on punishment behavior and the neural correlates of third-party punishment (TPP). The Marois lab and others have extensively studied the brain processes supporting third-party punishment (e.g. Buckholtz et al., 2008; Buckholtz et al., 2015; Ginther et al., 2016; Treadway et al., 2014), the proxy mechanism for legal decision-making. This expertise provides the theoretical and experimental groundwork on which we can launch the present study. Below we summarize the major findings from this line of research before elaborating on how we will leverage them to address the present aims.

Buckholtz and Marois (2012) suggest that TPP is supported by domain-general cognitive processes which process information relevant to punishment (i.e. harm, mental state, context) and integrate this information in order to reach a punishment decision. The integration of harm and mental state has consistently been found to predict TPP (Alter, Kernochan, & Darley, 2007; Carlsmith, Darley, & Robinson, 2002; Cushman, 2008; Ginther et al., 2016; Treadway et al., 2014). Evidence from imaging studies suggests that the processing of these two factors is initially distinct (Ginther et al., 2016; Treadway et al., 2014). The harm caused to a victim is associated with increased arousal in the amygdala (Buckholtz & Marois, 2012; Treadway et al., 2014; Ginther et al., 2016) as well as regions associated with perceptions of others' pain including the posterior insula, inferior parietal lob, and the orbitofrontal cortex (Ginther et al., 2016). By contrast, the evaluation of an offender's mental state shows increased activation in areas involved in mentalizing and engaging in Theory of Mind, especially the temporoparietal junction (TPJ).

Notably, the TPJ shows greater activation for harmful acts when the offender had diminished responsibility (Belluci et al., 2016; Buckholtz et al., 2008; Ginther et al., 2016; Treadway et al., 2014), and for reckless and negligent acts compared to clearly purposeful or blameless behavior. These findings suggest that TPJ activity is dependent on the amount of effort required to infer mental state rather than scaling linearly with the offender's culpability (Ginther et al., 2016). The use of TMS to disrupt activity in the right TPJ reduced the influence of the perpetrator's intent on participants' decisions, demonstrating a causal role of this region in engaging in theory of mind to assess intent of others during punishment decision-making (Young et al., 2010). Furthermore, increased activity in the TPJ for blameless harmful acts led to greater input from TPJ to the dorsal anterior cingulate cortex (dACC), which in turn engaged in top-down signaling of the amygdala to gate the affective response to the harm information, presumably decreasing punishment for blameless acts by attenuation of the emotional response (Treadway et al., 2014).

As mentioned above, the integration of harm and mental state is a strong predictor of TPP, as we do not engage in much punishment for blameless harmful acts or for mental states that are not harmful (Cushman, 2008; Alter et al., 2007; Ginther et al., 2016). Integration of harm and mental state has been observed in a selective number of brain regions, most notably the bilateral amygdala, medial prefrontal cortex (mPFC), right dorsolateral prefrontal cortex (DLPFC), and the posterior cingulate cortex (PCC). The bilateral amygdala in particular shows an activity pattern that reflects the interaction of harm and mental state in punishment response (Ginther et al., 2016). Finally, the decision phase – i.e. when participants decide whether and how much to punish the suspected wrongdoer – is associated with activity in rDLPFC, left ventrolateral prefrontal cortex (lVLPFC), and bilateral inferior fusiform gyrus (IFG) (Buckholtz et al., 2008; Ginther et al., 2016). The rDLPFC in particular shows a robust association with punishment decisions, though activity

does not correlate with punishment severity (Buckholtz et al., 2008; Buckholtz et al., 2015; Ginther et al., 2016). Together these studies highlight the importance of harm, intent, and their interaction in driving TPP neural network.

Here we assess the influence of character evidence (negative or neutral) and a subsequent legal admissibility instruction regarding that evidence (May Consider or Must Disregard) on TPP decisions as well as on the brain mechanisms involved in TPP described above. We adapted Ginther et al's (2016) paradigm – which allows for the isolation of the neural correlates of each phase of TPP decision making while parametrically manipulating the severity of the harm caused and the mental state of the perpetrator – by adding character evidence presentation and legal instruction information into the experimental paradigm while participants underwent brain scanning.

Methods

Participants

Twenty-five individuals (16 females; Mean age=23.24, range=18-28 with normal or corrected-to-normal vision completed the task. Participants were paid $50 for the two-hour session. Three participants were not included in the analysis due to excessive motion during scanning (>3 mm translation or 3 degrees of rotation), leaving 22 participants included in the analysis. All participants provided informed consent and the experimental protocol was approved by the Vanderbilt University Review Board.

Experimental Design

Participants made TPP decisions in response to hypothetical scenarios describing the actions of fictional protagonists. The task had a 2 (Character evidence: negative or neutral) x 4 (Mental State: blameless, negligent, reckless, purposeful) x 5 (Harm: none, minimal, moderate, life altering, death) x 2 (Instruction: "May Consider" or "Must Disregard") design. All variables were manipulated within-participants, and each participant completed 5 runs of 16 trials each, which allowed them to view one trial of each cell (80 trials total) via pseudorandom assignment. The runs lasted between 15 and 17 minutes depending on participant response time during the character and punishment stages.

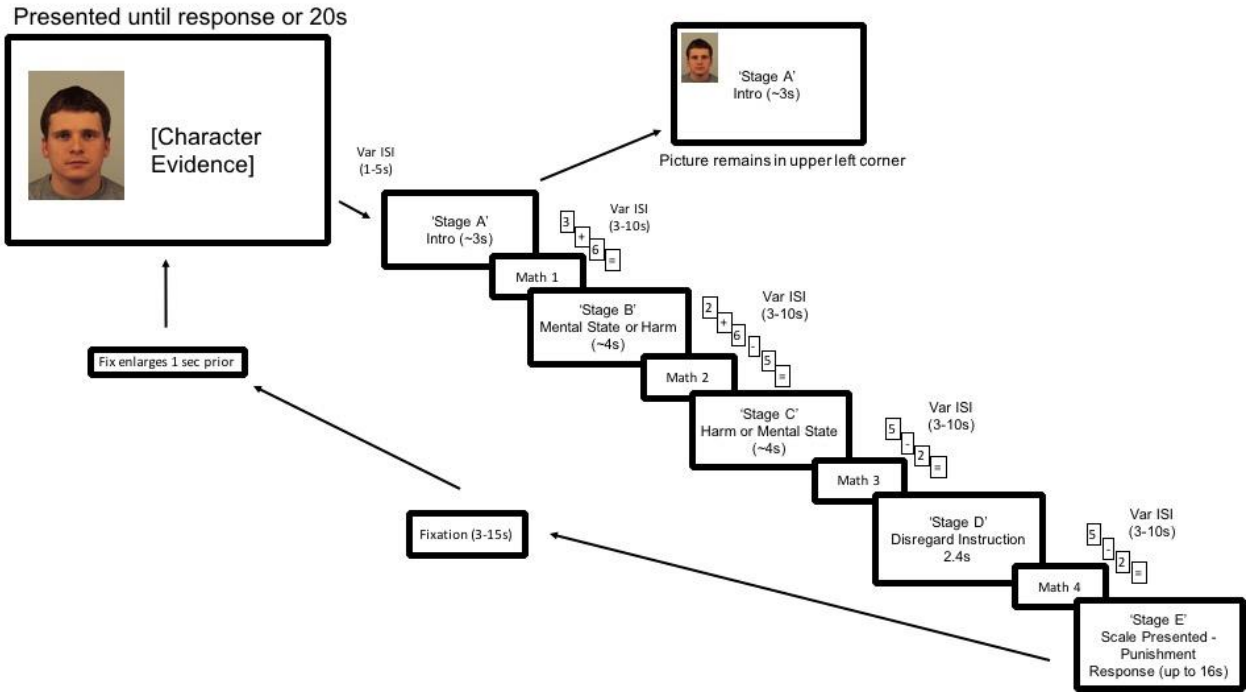

Figure 1. Trial for the behavioral paradigm participants completed during the fMRI scan

Behavioral Paradigm

We used the same behavioral paradigm employed previously by Ginther et al. (2016) with the addition of 1) a character evidence stage and 2) an instruction stage in regards to that evidence.

Figure 1 presents the trial design. Each trial began with the presentation of the character evidence; an image of the protagonist appeared on the screen along with a paragraph that presented either negative or neutral information about that individual (described below). The protagonist was always a male, and the image and name of the protagonist were randomly selected and were unique for each trial. This character information remained on the screen for 20 seconds or until the participant pressed a key to continue after reading the text. The trial scenario then began using the same parameters as in Ginther et al (2016). Each scenario contained three sentences which were presented separately: Stage A presented an introductory sentence describing the context in which the protagonist acted, while stages B and C described the harm or mental state, with the order of presentation for harm and mental state information randomized across trials. Following the scenario, participants received a legal instruction (Stage D) that told them whether they could consider the character information in their decision ("May Consider") or if they had to disregard that information and make their punishment decision using only the scenario ("Must Disregard"). The instruction appeared on the screen for 2.4 seconds. The final stage of the trial was the punishment decision (Stage E). Participants made a punishment decision on a 0 to 9 scale (described below) using button boxes within the scanner. They had up to 16 seconds to respond before the task automatically continued. An ITI was drawn from a decaying exponential distribution from 3 to 15 seconds, with the fixation square enlarging 1 second prior to the next trial to alert participants.

Ginther et al. (2016) used additional components within each trial to better isolate the time during which participants processed each stage of the task (i.e. mental state/harm presentation, integration, decision phase). We applied these same parameters for the present experiment. Specifically, the ISIs were filled with a secondary math task that spanned the length of each ISI.

Each math problem started 200 ms after each stage ended and included a series of addition or subtraction operations on integers between 1 and 9, with a solution between 0 and 9. The purpose of this task was to ensure that participants' processing of each stage was constrained to its presentation time and to prevent participants from thinking about their punishment response during the ISIs. The length of each ISI was randomly selected from an exponentially decaying distribution of 3-10 seconds. Additionally, each sentence of the scenario was presented as a rapid serial visual presentation (RSVP) rather than in paragraph form, with the words of the sentence presented sequentially at the center of the screen at the rate of 6 words per second. This also ensured that participants were only to processing the information presented during that specific Stage and not taking additional time to think about previous stages. We further randomly presented one of several available punishment scales during the punishment decision (Stage E) so that participants were forced to delay their punishment decision until they reached that stage.

Materials

Participants read either negative or neutral character evidence at the beginning of each trial. All character information was presented as a paragraph roughly 20 words in length, with 50 possible fact patterns per character type. Negative character evidence described prior violent or criminal acts that the protagonist had engaged in. Neutral character evidence provided information about the individual that described habits or facts about their life. Examples of each are provided below:

> <u>Negative Character Evidence-</u>
>
> In college [NAME] was convicted of sexual assault and battery towards his girlfriend, though he was released from prison over 10 years ago.
>
> [NAME] has been known to act violently towards animals, he tends to take out his frustrations by beating his dog.
>
> [NAME] was arrested 4 years ago for shooting a man in the leg and foot during a drug deal gone wrong.

Neutral Character Evidence-

[NAME] enjoys playing action video games with his friends to relax after he's had a long day at work.

[NAME] is very allergic to tree nuts and gets bright red hives if he is accidentally exposed to them.

[Name] drinks two large cups of coffee at home each morning to energize himself before he leaves for the day.

The scenarios describing the protagonist's actions parametrically manipulated the mental state of the protagonist and the severity of the harm they caused. We manipulated the mental state level using four of the five Model Penal Code categories; in increasing order of intentionality- purposeful, reckless, negligent, blameless (Ginther et al., 2014; Shen et al., 2011). The harm ranged from none, minimal, moderate, permanently life altering, and death. There were 80 different scenarios, each describing a different event involving the protagonist with a unique set of contextual facts (used previously in Ginther et al., 2016, Ginther et al., 2014, and Ginther, Hartsough, & Marois, under review). There were 16 scenarios describing each level of harm (16 scenarios x 5 levels of harm = 80 scenarios). Each scenario had a variation for each of the four mental state levels, and could be presented with either the harm or the mental state information first. Tables 1 and 2 present a sample scenario with each possible mental state variation with mental state presented first and harm presented first, respectively.

Illustrative Scenario (Planks & Bikes): Four Potential "Mental-State First" Variations.

Introductory Sentence			
John is hauling planks to his cabin because he is in the middle of doing carpentry work on his house, which abuts a public mountain bike trail.			
Mental State Sentence			
Purposeful Mental State	Reckless Mental State	Negligent Mental State	Blameless Mental State
Angry with the mountain bikers for making too much noise when biking past his house, John desires to injure some bikers by dropping planks on their trail so that they would hit them.	John drops some planks onto the trail without retrieving them because he's in a rush, even though he is aware there is a substantial risk bikers will hit them and be injured.	While John is carrying planks to his workshop in order to begin building new steps for his house, he drops some of the wood planks onto the bike trail without even noticing.	While John is carefully carrying some planks from his shed to the backyard, an unexpectedly strong gust of wind causes John to inadvertently drop several planks, despite his best efforts not to.
Harm Sentence			
Soon after John drops the planks, two bikers pass by and they hit the planks, which causes them to flip over their handlebars and one of the bikers suffers serious injuries as a result.			

Table 1. Sample scenario for a moderate level of harm at each mental state level, with mental state presented first followed by harm

Illustrative Scenario (Planks & Bikes): Four Potential "Harm First" Variations.

Introductory Sentence			
John is hauling planks to his cabin because he is in the middle of doing carpentry work on his house, which abuts a public mountain bike trail.			
Harm Sentence			
Soon after John crosses the trail, two bikers pass by and they hit planks that John dropped onto the trail, which causes them to flip over their handlebars and one of the bikers suffers serious injuries as a result.			
Mental State Sentence			
Purposeful Mental State	Reckless Mental State	Negligent Mental State	Blameless Mental State
Angry with the mountain bikers for making too much noise when biking past his house, John had desired to injure some bikers by dropping planks on the trail so that they would hit them.	John had dropped some planks onto the trail without retrieving them because he was in a rush, even though he was aware there was a substantial risk some bikers would hit them and be injured.	While John was carrying planks to his workshop in order to begin building new steps for his house, he had dropped some of the wood planks onto the bike trail without even noticing.	While John was carefully carrying planks from his shed to the backyard, he slipped on some mud, which caused him to unknowingly drop several planks, despite his best efforts not to.

Table 2. Sample scenario for a moderate level of harm at each mental state level, with harm presented first followed by mental state

The legal instruction stage included either "May Consider" in green text, or "Must Disregard" in red text. Participants were told that for the "May Consider" trials they could incorporate the character evidence into their punishment decision, while for the "Must Disregard" trials they were told they should put the character evidence out of their mind and use only the scenario information in the punishment decision.

Statistical Analysis: Behavioral Data

We conducted a mixed model repeated measures ANOVA to assess the effect of mental state, harm, mental state x harm, character, instruction, and character x instruction on punishment decisions. Subject was included as a random variable. For individual difference analyses, we performed a regression analysis for each individual subject to obtain beta values for how strongly subjects weighed the character evidence, admissibility instruction, and their interaction, using effect coding to obtain the main effects for both variables as they are categorical.

fMRI Acquisition

All fMRI scans were acquired using a 7T Philips Achieva scanner at the Vanderbilt University Institute of Imaging Science. Low- and high- resolution structural scans were first acquired using conventional parameters. Function (T2* weighted) images were acquired using a gradient-echo echoplanar imaging (EPI) pulse sequence with the following parameters: TR 1000 ms, TE 40 ms, flip angle 79°, FOV 240 X 111 X 240 mm, with 25 axial slices (3.0mm, 1.5 mm gap) oriented parallel to the AC-PC line and collected in an ascending interleaved pattern.

Statistical Analysis: fMRI Data

Analysis was conducted using Brain Voyager QX 2.8 in conjunction with custom MATLAB software, using the same approach as in Ginther et al. (2016). All functional images were preprocessed using slice timing correction, 3D motion correction, linear trend removal (1/128 Hz), temporal high pass filtering, and spatial smoothing with a 6 mm Gaussian kernel as implemented through Brain Voyager software. Each participants' functional data were aligned with their anatomical volumes and transformed into standardized Talairach space.

We created design matrices for each participant by convolving the task events with a canonical hemodynamic response function. For the task events, the presentation of each stage of a scenario was modeled as a boxcar function spanning the duration of the stage. We also inserted 6 estimated motion parameters (X, Y, and Z translation and rotation) as nuisance regressors into each design matrix.

For our first-level analysis, we created GLMs for each subject's data to model different stages of the task. We used the same GLMs as in Ginther et al. to identify brain regions involved in processing context, harm and mental state information, those involved in integration of harm and mental state, and those involved in punishment decision and response (see Supplementary Materials Section 2). To assess the processes involved in the evaluation of character evidence, we created three GLMs with different regressors for negative and neutral character evidence modeled for the character evidence presentation stage (Stage A), instruction stage (Stage D), and decision stage (Stage E). We similarly created two GLMs to assess the processes involved in the evaluation of the admissibility instruction by modeling different regressors for the May Consider and Must Disregard instructions at Stage D and Stage E. To model the cognitive processes recruited by the different task stages, regardless of the information presented at the stage, we modeled each stage of the task as well as the ISI math task. All GLMs were created using z-transformed time course data, as in Ginther et al., (2016).

Second-order random-effects were conducted on each subject's beta weights. We apply a False Discovery Rate (FDR) to control for multiple comparisons ($q<0.05$ with $c(V)=1$) and apply a 10 voxel cluster size minimum. For conjunction analyses we applied a minimum test statistic (Ginther et al., 2016; Nichols et al., 2005). When we perform post hoc analyses on regions identified via whole-brain analyses, we control for multiple comparisons again using a FRD

threshold of $q<0.05$. To test for effects within ROIs identified in Ginther et al. (2016) we created 8mm spheres around the peak voxel for each ROI. To identify sub-clusters within large regions of activation for the present study, local maxima were identified using a higher-values-first watershed search algorithm implemented via BrainVoyager QX and NeuroElf. The algorithm works as follows: all voxel values within a given cluster are sorted from highest to lowest, with the highest voxel coordinate marked as "sub-cluster 1". Each voxel in turn (highest to lowest) is tested for whether it is linked to an existing sub-cluster by determining the minimum distance to voxels that are already marked. All voxels directly connected without interruption are therefore marked as belonging to the same sub-cluster; voxels connected to several sub-clusters go to the sub-cluster with the highest value (NeuroElf; http://neuroelf.net). We then created ROIs as 8mm spheres around each local maximum identified.

Behavioral Results

Figure 2 shows mean punishment ratings as a function of harm, mental state, character evidence, and instruction. A repeated measures ANOVA for the punishment decisions found significant main effects for mental state ($F(3, 63) = 85.70$, $p<.001$), harm ($F(4, 84) = 132.58$, $p<.001$), and instruction ($F(1, 21) = 8.66$, $p=0.004$). Punishment increased for both culpable mental states and more severe harms, and was greater when participants were told they could consider the character evidence versus being instructed to disregard it. The interaction term for mental state x harm was also significant ($F(12, 252) = 7.70$, $p<.001$) with greater punishment for more severe culpable harms, as we expected given the robust nature of this interaction in previous third-party punishment studies (see e.g. Ginther et al., 2016). There was no main effect of character evidence ($F(1, 21) = 1.48$, $p=0.35$) and no interaction between character evidence and instruction ($F(1, 21)$

= 0.84, p=0.36). We had expected both an effect of character and an interaction between character evidence and instruction based on a pilot study conducted outside of the scanner (see Supplementary Materials Section 1). However, while the design of this pilot study was otherwise nearly identical to that of the fMRI behavioral task, a potential key difference was that the character evidence was presented a second time on screen during the punishment decision (Fig. 1, Stage E; but see Behavioral Control Experiment below); this was removed for the fMRI experiment in order to confound the isolation of punishment decision-related activity with character evidence.

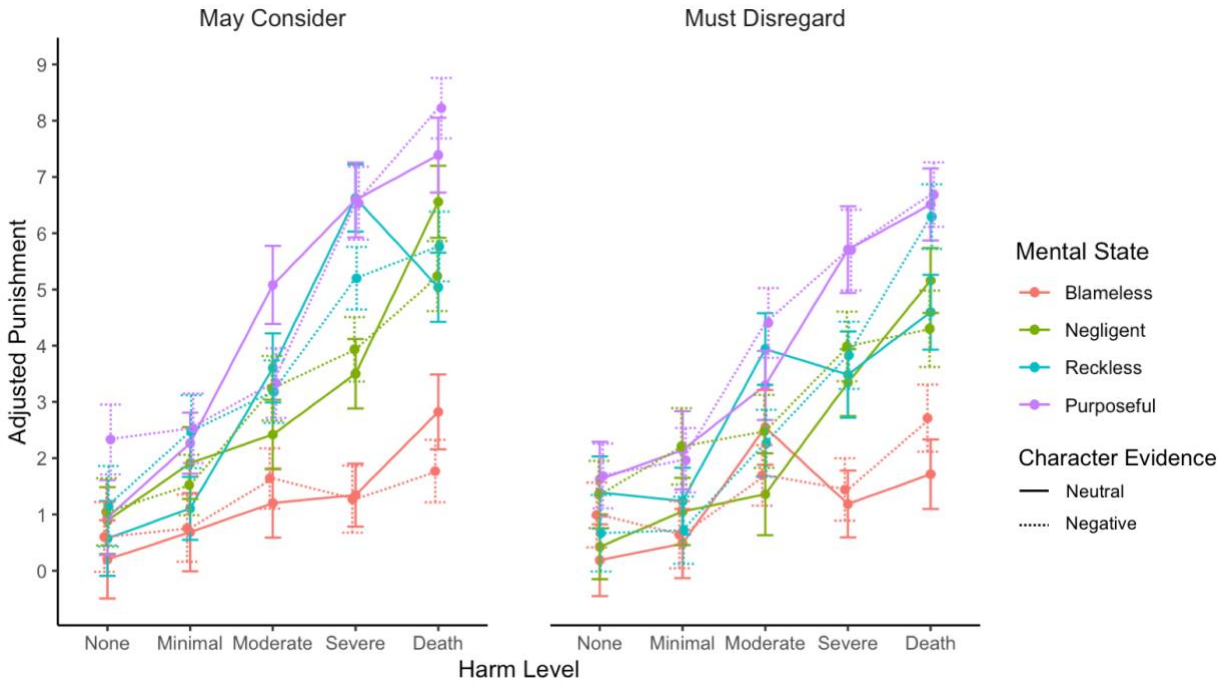

Figure 2. Standardized mean punishment amount +/-SE by harm (x-axis), mental state (colored legend), character evidence (solid line neutral, dotted line negative), and instruction ("May Consider" on the left, "Must Disregard" on the right).

Behavioral Control Experiment

The above behavioral results demonstrated increased punishment behavior when participants were told they could consider the character evidence compared to when they received

a disregard instruction. What remains unclear however, is how the instruction is influencing punishment, particularly in the absence of an effect of character evidence. Specifically, it is possible that regardless of the character evidence presented: 1) the "May Consider" instruction leads to greater punishment; 2) the "Must Disregard" instruction results in less punishment; or 3) that both these effects occur. Unfortunately, the behavioral data above cannot distinguish between these possibilities, nor is the previous literature categorical on this issue. As described in the Introduction, prior research on the application of legal disregard instructions shows that receiving a disregard instruction does not necessarily lead individuals to punish in the same way they would had they truly never heard the information to begin with (Carretta & Moreland, 1983; Casper et al., 1989; Edwards & Bryan, 1997; Lieberman & Arndt, 2000; Sue, Smith, & Caldwell, 1973; Tanford & Cox, 1988). Further, there is some evidence to suggest that the disregard instruction itself may influence punishment outcomes (Broeder, 1959; Lieberman & Arndt, 200; Pickel, 1995; Thompson et al., 1981).

To better understand how instructions affected punishment, we conducted a behavioral control experiment in which we included a condition where participants are presented with character evidence but received no instruction as to whether they should apply that information, as well as a control condition in which participants did not receive any character information (and therefore no instruction). These conditions will help elucidating the processes involved in the application of the disregard instruction and thus help interpreting the behavioral results of the fMRI experiment.

Methods

We recruited 24 participants (18 females, Mean age=22.33) to complete the task in the lab. Participants received $18 for completing the 1.5-hour study. All participants provided informed consent and the experimental protocol was approved by the Vanderbilt University Review Board.

The task employed a 2 (character evidence: negative or neutral) x 4 (mental state: blameless, negligent, reckless, purposeful) x 5 (harm: none, minimal, moderate, life altering, death) x 3 (instruction: "May Consider", "Must Disregard", or no instruction) + control trials (no character evidence and no instruction) design. All variables were manipulated within-participants, and each participant completed 5 blocks of 16 trials each (80 trials total) as in the fMRI behavioral task. We used pseudorandom assignment to ensure relatively equal sample sizes between all possible cells (140 cells total: (2 x 4 x 5 x 3) + (4 x 5 control trials)).

Trials that include both character evidence and an instruction were identical to those in the fMRI behavioral task. For trials with no character evidence (control trials), a picture of the protagonist appeared as before but rather than viewing a paragraph of character evidence participants were simply prompted to "Press any key to continue" to continue to the scenario. For trials with no instruction (both the 'no instruction' condition and control trials), a fixation square appeared for the 2.4 seconds during Stage D rather than the instruction language. We did not include a jittered ITI for this behavioral experiment; the ITI was always 1 second to reduce the total experiment time. We kept the math task ISIs in order to keep the cognitive load/distractions between stages consistent with the fMRI behavioral task.

Behavioral Control Results

Figure 3 presents mean punishment ratings by harm, mental state, character evidence, and instruction. Since we did not have a fully crossed design due to the inclusion of the control

condition, we conducted a repeated measures ANOVA for harm x mental state plus condition (7 levels: 2 character levels x 3 instruction levels, plus control). The results show main effects of harm ($F(4, 92) = 445.01$, $p<.001$), mental state ($F(3, 69) = 235.15$, $p<.001$), and condition ($F(6, 138) = 7.29$, $p<.001$), as well as an interaction between harm and mental state ($F(12, 276) = 24.5$, $p<.001$). All these effects are consistent with the fMRI behavioral data described above.

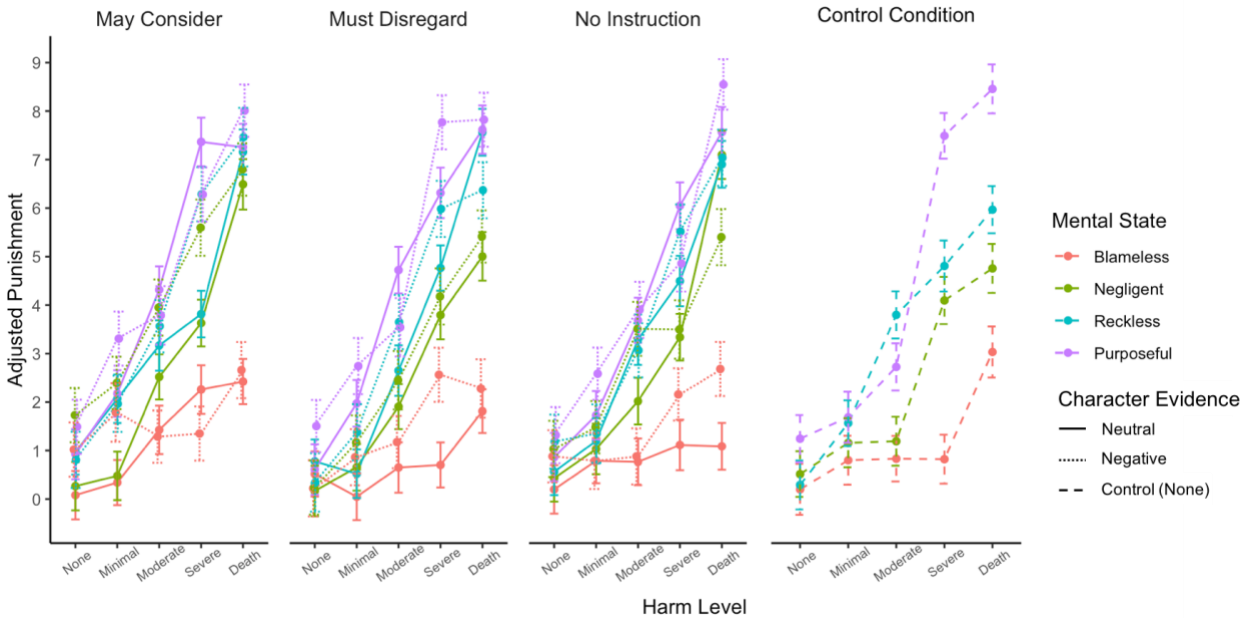

Figure 3. Standardized mean punishment amount +/-SE by harm (x-axis), mental state (colored legend), character evidence (solid line neutral, dotted line negative, dashed line control/no evidence), and instruction (plot panels; control condition had no character evidence or instruction).

Figure 4 shows mean punishment ratings by character evidence, instruction, and the control condition, collapsed across mental state and harm. Post-hoc pairwise comparisons between conditions with Bonferroni-adjusted p values found that punishment in the Negative Character-May Consider condition was significantly greater than in Neutral Character-Must Disregard (adjusted $p=0.003$), Neutral Character-No Instruction (adjusted $p=0.01$), and the Control conditions (adjusted $p=0.04$); no other comparisons were significant.

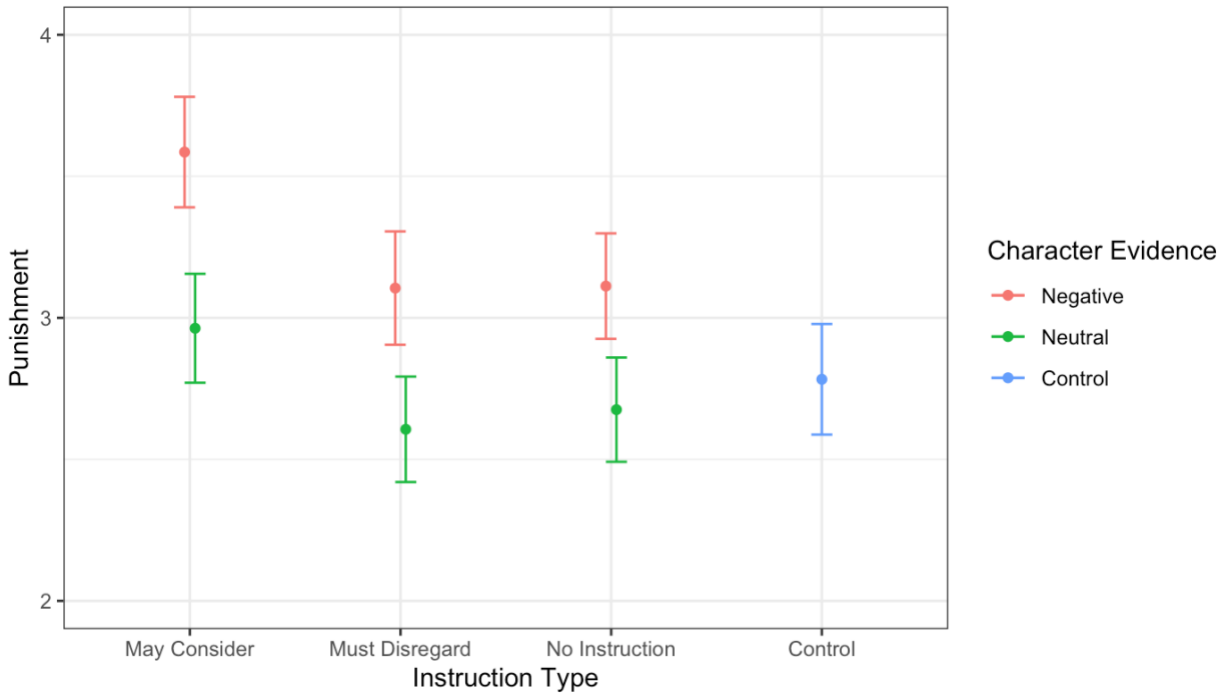

Figure 4. Standardized mean punishment amount +/-SE by character evidence (colors) and instruction (x-axis) plus the control condition (no character evidence or instruction). While the punishment scale went from 0 to 9, we have limited the y-axis between 2 and 4 to better visualize the results.

Furthermore, when we performed a repeated measures ANOVA after excluding the control condition (in order to test the main effects of character evidence, instruction, and their interaction) we found main effects of harm ($F(4, 92) = 376.07$, $p<.001$), mental state ($F(3, 69) = 199.78$), character evidence ($F(1, 23) = 25.73$, $p<.001$), and instruction ($F(2, 46) = 6.80$, $p=0.001$). Post-hoc pairwise comparisons between instructions with Fisher's LSD-adjusted p values found that the effect of instruction was due to a greater punishment for May Consider compared to both Must Disregard (adjusted $p=0.02$) and No Instruction (adjusted $p=0.04$), with no difference between Must Disregard and No Instruction (adjusted $p=0.81$). There was also an interaction between harm and mental state ($F(12, 276) = 21.23$, $p<.001$), but no interaction between character evidence and instruction ($F(2, 46) = 0.47$, $p=0.63$).

The behavioral control experiment revealed several key findings. First, we observed an effect of character evidence that was not seen in the fMRI behavioral data, which suggests that this information can affect participant decisions within the overall paradigm. We believe that the lack of a behavioral effect for character in the fMRI experiment may be due to the increased demands of completing the task within the 7T scanner setting. Anecdotally, participants reported that the fMRI task was challenging and placed a high demand on working memory, particularly since they had to make responses to the math problems within each ISI using the scanner button boxes, which require associating each finger with a memorized number versus simply pressing the corresponding number key on a keyboard in the behavioral control experiment. It may therefore be that the fMRI task was so cognitively taxing and stressful for the participants that they struggled to maintain or retrieve the character evidence information during each trial. More importantly, the control experiment demonstrated that the Must Disregard instruction did not impact participants' punishment decisions, as punishment for this instruction was not different compared to when participants received No Instruction or for the Control condition for both negative and neutral character evidence. Conversely, the May Consider instruction increased punishment decisions, and while this effect was greatest for negative character evidence, it was also true relative to the Must Disregard and No Instruction conditions when controlling for the effect of character evidence. These results indicate that the 'May Consider' Instruction increases third-party punishment whereas the 'Must Disregard' instruction does not affect punishment above and beyond having no instructions or no character evidence, and this result is obtained irrespective of the nature of the character evidence (negative or neutral). These results have important implications not only for the use of Disregard Instructions in the legal system, but also more pertinently for the present fMRI experiment, as they suggest that the May Consider instructions encourage participants to punish

more severely than the Must Disregard Instructions might blunt or suppress punishment behavior. This account explains well how the Disregard instructions had a significant effect on punishment behavior even when in the absence of an effect of character (i.e. no difference between negative and neutral) in the fMRI experiment.

fMRI Results

We first repeated the analyses of Ginther et al. (2016) to identify brain regions involved in processing context, harm and mental state information, those involved in integration of harm and mental state, and those involved in punishment decision and response. These analyses were broadly consistent with those of this previous study, (see Supplementary Materials Section 2).

The subsequent fMRI analyses sought to address the following questions: 1) What brain regions are involved in the evaluation of character evidence and admissibility instructions related to that evidence? 2) How does character evidence influence brain regions previously identified as part of the TPP neural network (esp. Ginther et al., 2016), as well as regions identified by question 1? And 3) How do admissibility instructions modulate activity in the brain regions previously identified as being part of the TPP neural network (see above), as well as regions identified by question 1? To address these questions, we first carried out voxel-based SPM analyses to identify brain regions that may be associated with each of them. We then probed these ROIs and Ginther et al (2016) ROIs for character evidence and Disregard Instruction effects. Finally, we applied an individuals' difference analysis to identify brain regions that correlated with variance in subjects' behavioral performance.

Character Evidence

BOLD amplitude analysis

To identify regions involved in the evaluation of character evidence we first performed voxel-based, whole-brain contrasts of Negative versus Neutral character evidence during the presentation of the character evidence (Stage A), as well as during the presentation of the instruction (Stage D) and during punishment decisions (Stage E); none of these contrasts yielded significant activation. This is consistent with the lack of a group-level behavioral effect of character evidence for these participants. We then identified regions involved in processing character evidence regardless of valence by performing a conjunction contrast between Stage A (presentation of character evidence) and all other task stages. Specifically, we used the GLM that modeled each stage of the task to perform a conjunction analysis of Stage A compared with each of the other task conditions, namely, mental state and harm evaluation, instruction evaluation, decision stage, and the ISI math task (same approach used to assess areas involved in decision stage in Ginther et al. 2016). The only region to show significant activation was the right occipital gyrus, most likely owing to the presentation of the large picture of the trial protagonist at this stage (see Fig. 1). An open contrast during character presentation (Stage A) resulted in a number of active regions whose local maxima were identified using a higher-values-first watershed search algorithm implemented via BrainVoyager QX and NeuroElf, as previous methods for identifying local maxima (e.g. Monte Carlo) showed high false positives (Eklund, 2016). Table S3 lists the four large clusters identified for the open contrast at character presentation as well as the identified sub-clusters for each.

To further explore for any trace of character evidence in the brain, we first compared mean beta activity for negative versus neutral character evidence within the groups of ROIs identified in Ginther et al. as well as in the Conjunction contrast ROIs identified for character evidence as

described above. The only regions to show an effect of character evidence after correcting for multiple comparisons (again using a FDR threshold of $q < 0.05$) were the LTPJ identified in Ginther for the MS>Harm contrast ($t=3.84$, adjusted $p=0.01$) and the LSTS identified for the present data for the same contrast ($t=3.003$, adjusted $p=0.03$), both at the time of character evidence presentation. These regions are associated with a Theory of Mind network and it makes sense that we would see activation when they are evaluating the protagonist. Importantly, we observed this effect of character evidence in these regions only at the time of character evidence presentation; the fact that this effect did not persist at later stages is consistent with the lack of a group-level behavioral effect of character evidence on participants' ultimate punishment decisions.

Individual Difference Analyses

Closer examination of the behavioral and fMRI data revealed a large variance in performance and activity pattern. Such variance may obscure group-based effects. Another means to assess the effect of variables onto behavior and neural activity is by means of individual differences analysis. The logic of this analysis is to identify brain areas whose activity may account for the variability in behavioral performance across subjects, thus implicating such brain areas in that performance. Turning this individual differences' analysis to character evidence, we sought to identify brain regions involved in character evidence by correlating the beta weight difference (of punishment) between negative and neutral character evidence to the beta weight difference in BOLD activity between negative and neutral character evidence at the time of evidence presentation. We did not observe any effect of individual differences for character evidence in any Conjunction contrast ROIs or Ginther ROIs. The same held true for examination of these ROIs at the instruction or decision stages.

Admissibility Instructions

We used similar analytical approaches to identify regions involved in the evaluation of the admissibility instruction. First, we performed whole brain contrasts of the May Consider vs Must Disregard instructions during the presentation of the instruction (Stage D) and during punishment decisions (Stage E); none of these contrasts yielded significant activation that survived correction for multiple comparisons. We then identified regions involved in evaluating the admissibility instruction regardless of instruction type by performing a conjunction contrast between Stage D (presentation of the instruction) and all other task stages. This resulted in a number of active regions (Table S5 lists the 8 large clusters identified in the instruction > all other stages contrast and the identified sub-clusters for each).

BOLD amplitude analysis

To assess the effect of admissibility instruction on the neural correlates of TPP we similarly compared mean beta activity for May Consider versus Must Disregard instruction within the same groups of ROIs as above (i.e. both the ROIS listed in Table 5 and the Ginther et al., ROIs), at different stages of the task. We did not find any effect of instruction in any of these ROIs, despite a group-level behavioral effect observed in the behavioral results.

Individual Differences Analysis

We then applied this behavioral differences analysis to Disregard Instructions. We correlated subjects' behavioral instruction beta weight with the difference in neural activity they demonstrated for the May Consider > Must Disregard contrast for each ROI with an FDR ($q<0.05$) correction for multiple comparisons.

Within the Conjunction contrast ROIs identified during Instruction presentation (Stage D), none showed activity correlation with subjects' behavioral instruction beta weight. The same held

for the Ginther ROIs. However, three Instruction conjunction contrast ROIs showed a significant correlation during the punishment decision stage (Stage E); right inferior frontal gyrus (r=0.64, p=0.001; Brodmann area 44; Fig. 5A) left medial frontal gyrus (r=0.60, p=0.003; Fig. 5B), and left superior frontal gyrus (r=0.67, p=0.001; Fig. 5C). We also observed similar correlations in the mPFC (r=0.57, p=0.006; Fig. 5D) and right middle occiptal gyrus (r=0.59, p=0.004; Fig. 5E) of the Ginther ROIs. In each of these regions, participants showing the strongest behavioral effect of instruction (i.e. showing greater punishment for May Consider vs Must Disregard) had a greater difference in activation for the May Consider > Must Disregard contrast, suggesting that these areas may play a role in the evaluation of the admissibility instruction at the time of punishment decision.

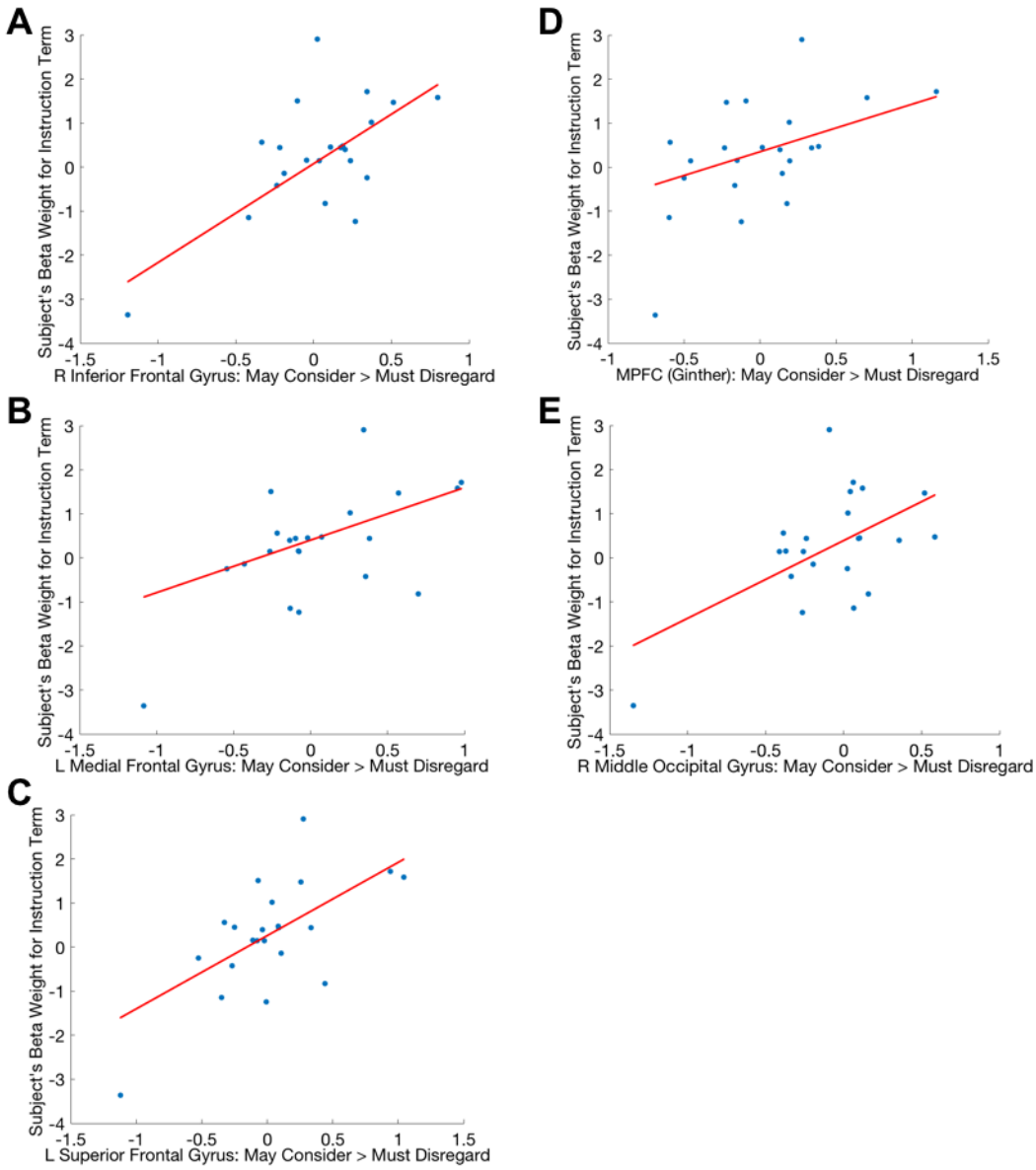

Figure 5. Correlation plots for areas showing significant correlation for individual behavioral beta weight for instruction term and May Consider>Must Disregard contrast at time of punishment decision. A. Right Inferior Frontal Gyrus; B. Left Medial Frontal Gyrus; C. Left Superior Frontal Gyrus; D. mPFC from Ginther et al.; E. Right Middle Occipital Gyrus from Ginther et al.

Discussion

In this study we sought to evaluate how character evidence and admissibility instructions related to that evidence influence the behavior and neural correlates of TPP. We added stages presenting these elements to the task design used in Ginther et al. (2016), which identified key

brain regions involved in the assessment of harm and mental state information, the integration of this information, and the subsequent punishment decision. Our primary findings were that the admissibility instruction influenced punishment decisions even in the absence of an effect of character evidence, and that participants who were more responsive to differences for the instruction showed corresponding activation in several brain regions.

Our behavioral analyses from the fMRI experiment found a main effect of admissibility instruction provided to participants but not of character evidence. Since we found an effect of character evidence using the same general behavioral paradigm in a pilot study as well as in the behavioral control study, it seems that the lack of a group-level effect of character evidence for our fMRI participants may be due to additional challenges related to performing the task in the scanner. Participating in an fMRI experiment at 7T can be stressful, and responding using the scanner button boxes puts additional strain on participants' working memory, as they have to maintain a mapping of the numbers associated with each finger while trying to rapidly evaluate the information and respond to math problems between each stage of the task. In correspondence to the behavioral results, we found little neural trace of character evidence in the brain imaging analyses, save for greater activity (beta weights) for negative character than neutral character evidence in the left TPJ and STS, and only during character presentation. This activity probably reflects the greater mentalizing demands in evaluating norm-breaking behavior, a role previously associated with these brain regions (e.g. Ginther et al., 2016; Young et al., 2010). Furthermore, the fact that this activity trace did not persist beyond the character evidence presentation stage is consistent with the lack of a differential effect of character evidence at punishment, perhaps because the participants could not maintain that information throughtout the duration of the cognitively demanding trials. If it is true, this hypothesis would suggest that we should observe a

neural trace of character evidence in the TPJ (and maybe other area) at the decision phase in a paradigm (like the pilot study) that exhibits a behavioral effect of character evidence in punishment decisions.

It is interesting that even in the absence of a behavioral effect of character evidence, we observed a behavioral effect of the instructions stating the admissibility of that evidence, with greater punishment for the May Consider versus Must Disregard instruction. What we could not determine from the behavioral data of the fMRI expeirment alone was whether this effect was due to the influence of the May Consider instruction, the Must Disregard instruction, or both. Our behavioral control study shed light on this issue by adding a condition in which participants receiving no instruction and another (control) condition that had neither character evidence nor an instruction. We found that punishment decisions were comparable for the Must Disregard instruction and the No-Instruction and Control conditions but greater for the May Consider instruction, even when controlling for character evidence. This suggests that the behavioral effect of instruction for our fMRI participants was due primarily to increased punishment in response to the May Consider instruction. Previous studies have found that while a disregard instruction may not eliminate all the bias introduced by inadmissible evidence, decisions typically are not as severe as when the evidence is deemed admissible (e.g. Carretta & Moreland, 1983; Kassin & Sommers, 1997), which is generally consistent with our behavioral control results. It is important to note however, that while past research has assessed the effect of negative evidence that has been deemed admissible or inadmissible in comparison to control conditions (no evidence/no instruction), to our knowledge previous work has not evaluated the equivalent of our neutral conditions. Given our finding that the May Consider instruction has an effect independent of an effect of character evidence, it may be worth considering how declaring evidence admissible may influence behavior

beyond the effect of allowing the evidence to be considered. Indeed, we surmise that the very act of approving the admission of character evidence may encourage jurors to render an unfavorable decision towards the defendant by implying that this evidence is relevant, or by inflating its importance, to the case at hand.

We found that individuals who showed a stronger behavioral effect of instruction (i.e. punish more for May Consider versus Must Disregard) show greater activation during the punishment decision stage for the May Consider > Must Disregard contrast in multiple areas of theprefrontal cortex; specifically, the right inferior frontal gyrus, left superior frontal gyrus, left medial frontal gyrus, and the mPFC. Activity in the right inferior frontal gyrus (Brodmann area 44) has been associated with motor inhibition (Bernal & Altman, 2006; Neef et al., 2016) and working memory (Fiebach, et al., 2005; Ranganath, Johnson, & D'Esposito, 2003). Interestingly, intentional remembering is associated with activity in left-lateralized regions of the PFC including the superior and medial frontal gyri (Rizio & Dennis, 2013). This may suggest that participants who show greater activation and punishment for the May Consider instruction are attempting to retrieve information from earlier in the trial when they are told they can use that information in their punishment decision. The mPFC region from Ginther et al. (2016) is thought to be involved in integrating information prior to the punishment decision (Buckholtz & Marois, 2012; Sporns et al., 2007). Together, increased activation in these areas in response to the May Consider > Must Disregard contrast among participants who punish more for May Consider is consistent with the idea that these individuals do make an effort to recall and integrate information into their decisions. It is less clear, however, what information is being recalled or integrated that may contribute to increased punishment.

We acknowledge several limiations to the present study. The first is that we did not observe the expected behavioral effect of character evidence for our fMRI participants, which limited our ability to assess the neural correlates of the evaluation of the evidence as well as our interpretations of the effects of instruction (both behaviorally and neurally). In that regard, it will be worth conducting a multivoxel pattern analysis (MVPA) to determine whether the different character evidence valences can be differentiated within regions found to activate for the character stages (and subsequent stages).The same analysis should also be carried out for the Instruction manipulation. The MVPA analysis could provide additional insight into the influence of these variables on the neural correlates of TPP, which was a primary motivation for the present chapter. Furthermore, as with our other experiments we acknowledge that we have limited external validity in regards to the application of these findings to a broader range of legal contexts and to actual jury decision-making. For this study specifically, we had participants provide a punishment rating rather than making a verdict decision; future research should expand on the present findings in other contexts not only because jurors typically do not make punishment decisions but also because previous research has found that disregard instructions may influence different types of legal decisions (i.e. verdicts, guilt ratings, perceived credibility of defendant) differently (e.g. Carretta & Moreland, 1983; Tanford & Cox, 1988). Nonetheless, our findings in this study again demonstrate the importance of instructions within the legal domain, and consistent with results throughout the thesis suggest that the instructions themselves may affect behavior in ways that are not accounted for in the legal system.

CHAPTER 7

GENERAL DISCUSSION

We conducted a series of experiments to assess the effect of legal burden of proof and disregard instructions on the behavioral and neural mechanisms of decision-making. Our findings show that legal instructions have complex and varied effects on decisions in ways that are dependent on context and the expertise of the decision-maker. As discussed below, this has implications not only for human psychology but also for understanding the processes involved in high stakes legal decision making.

In Chapters 2-5 we implemented a psychometric approach to estimate and compare decision parameters, which allowed us to assess the effect of context, expertise, and cost related to the application of burden of proof instructions (i.e. PoE, BaRD). This approach was advantageous because it provided multiple parameters that gave information about decision processes. Our findings show that people generally do not differentiate adequately between these standards (i.e. too stringent PoE, too lenient BaRD). We also consistently found that decisions were more conservative within the legal domain relative to non-legal domains (medical, scientific, general), and that this effect lead to the overly stringent application of the PoE standard in comparison to the prescribed threshold of just over 50% as well as compared to individuals' intuitive decision thresholds (IB).

Interestingly, we found that expertise affected the application of instructions in Chapter 3. Our findings suggest that individuals' intuitive decision thresholds are influenced by their experience, particularly when that expertise is associated with the application of standard decision

criteria in their field of expertise. While this influence led to greater differentiation of the PoE and BaRD standards for legal experts (who were consistent with the prescribed standards), scientific, medical, and humanities experts were generally more conservative than non-experts. As noted in Chapter 3, this may have implications in jury selection, as academics may adopt more stringent decision criteria than the general population. We also found that the increased stringency for decisions in a legal context extended to experts in the non-legal fields as well as laypeople, which may suggest that a lack of familiarity or expertise in the law contributes to the effect of legal domain as well as overly stringent application of the PoE standard.

Chapter 4 aimed at elucidating why PoE is consistently applied more stringently by laypeople than its intended standard and individuals' own belief. Our findings suggest that it is the use of the phrase "preponderance of the evidence" applied in a legal context that leads to this disparity in laypeople's application of PoE. Participants were able to apply the threshold correctly when using other equivalent instructions without the PoE phrase included, suggesting that the justice system should consider the adoption, for instance, of a "more likely true than not" instruction in place of the less familiar and potentially biasing "preponderance of the evidence".

Chapter 5 aimed at understanding why individuals' decision making is consistently more conservative in the legal domain than in other domains. Specifically, we tested the hypothesis that it is because they are implicitly weighing the potentially serious consequences (i.e. punishment) that could result from rendering legal decision. We found that decision costs have little impact on individuals' decisions, both within the legal domain as well as across other non-legal domains. This finding is consistent with the goals of the legal system with respect to the application of the burden of proof standards, as jurors are not supposed to consider potential sentencing when reaching a verdict for a defendant, and are meant to apply decision standards equitably across cases

(Laudan, 2003). Strikingly, decisions were more conservative for the legal domain than other societal domains regardless of the actual punitive cost, which suggests that it is the nature of the cost within a legal context (i.e. always a negative consequence to the defendant) that influences decisions rather than the absolute value of the cost. Given the high costs to both individuals and society described in the General Introduction, it may be reassuring to note that legal decisions are inherently conservative, though we note that for laypeople the application of the BaRD standard may still not be as conservative as intended.

In Chapter 6 we assessed the application of admissibility instructions to either consider or disregard potentially biasing character evidence. Our behavioral results indicate that these instructions influence third-party punishment decisions even in the absence of (fMRI experiment) or controlling for (behavioral control experiment) an effect of character evidence, and that this is due to the 'May Consider' instruction leading to greater punishment. This finding has implications for understanding how admissibility rulings influence juror behavior, as past research has focused on how a disregard instruction may or may not affect decisions, but has not explored the possibility that a ruling of admissible may itself affect decisions beyond just the consideration of the related evidence. Admitting evidence that has been challenged may also imply that the evidence is relevant or important which could cause individuals to weigh it more heavily. Our fMRI results found that participants who punished more under 'May Consider instructions (controlling for character evidence) exhibited greater activation under the same instructions in several prefrontal regions involved in working memory and intentional remembering, as well as in part of mPFC associated with information integration. We propose that these individuals are attempting to recall and integrate additional information into their punishment decision when told they can consider the

evidence, though it is not yet clear what information they may actually remember or use that contributes to greater punishment.

As a whole, the present body of work highlights the importance of instructions within the legal domain and reveals lines of faults in their intended applications, both by laypeople and experts. Furthermore, our studies isolate the sources of these faults or limitations. In turn, these findings have important implications for the use of these instructions in legal decision-making, for they may lead to the reformation of legal standards and evidentiary instructions for the purpose of striving for a fairer justice system.

REFERENCES

Alter, A. L., Kernochan, J., & Darley, J. M. (2007). Transgression Wrongfulness Outweighs its Harmfulness as a Determinant of Sentence Severity. *Law and Human Behavior, 31*. 319-335.

Arkes, H. R. & Mellers, B. A. (2002). Do juries meet our expectations? *Law and Human Behavior, 26*(6), 625-639.

Balliet, D. & Van Lange, P. A. M. (2013). Trust, Punishment, and Cooperation Across 18 Societies: A Meta-Analysis. *Psychological Science, 8*(4). 363-379.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K. M., Grafman, J., & Krueger, F. (2016). Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. *Social Neuroscience, 10*. 1-11.

Boettrich, S. & Starykh, S. (2019). Recent Trends in Securities Class Action Litigation: 2018 Full-Year Review. *NERA Economic Consulting.* NERA.com

Boland, M. & Lehmann, H. (2010). A new method for determining physician decision thresholds using empiric, uncertain recommendations. *BMC Medical Informatics and Decision Making, 10,* 20.

Bright, D. A. & Goodman-Delahunty, J. (2006). Gruesome Evidence and Emotion: Anger, Blame, and Jury Decision-Making. *Law and Human Behavior, 30*. 183-202.

Broeder, D. W. (1959). The University of Chicago Jury Project. *Nebraska Law Review, 38*(3). 744.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*. 930-940.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience, 15*(5). 655-661.

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From blame to punishment: Disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron, 87*(6). 1369-1380.

Bureau of Prisons. (2019). Annual determination of average cost of incarceration fee. *Federal Register, 84 FR 63891,* 63891.

Butler, A. & James, K. (2010). The neural correlates of attempting to suppress negative versus neutral memories. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 182-194.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why Do We Punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2). 284-299.

Carretta, T. R. & Moreland, R. L. (1983). The direct and indirect effects of inadmissible evidence. *Journal of Applied Social Psychology, 13*(4), 291-309.

Casper, J. D., Benedict, K., & Perry, J. L. (1989). Juror decision making, attitudes, and the hindsight bias. *Law and Human Behavior, 13*(3), 291-310.

Champion, D. J. (1989). Private counsels and public defenders: A look at weak cases, prior records, and leniency in plea bargaining. *Journal of Criminal Justice, 17*(4), 253-263.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3).

Cushman, F. (2008). Crime and Punishment. *Cognition, 108*. 353-380.

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of*

Ibadan Postgraduate Medicine, 6(1), 21-26.
Dhami, M. K., Lundrigan, S., & Mueller-Johnson, K. (2015). Instructions on reasonable doubt: Defining the standard of proof and the juror's task. *Psychology, Public Policy, and Law, 21*(2), 169-178.
Djulbegovic, B. et al. (2014). How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making, 14*, 47.
Edwards, K. & Bryan, T. S. (1997). Judgmental biases produced by instructions to disregard: The (paradoxical) case of emotional information. *PSPB, 23*(8), 849-864.
Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519-527.
Fehr, E. & Fishbacher, U. (2004). Social norms and human cooperation. *Evolution and Human Behavior, 25*. 63-87.
Fehr, E. & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415*. 136-140.
Fründ, I., Haenel, N., & Wichmann, F. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision, 11*(6), 1-19.
Gilchrist, J. M., Jerwood, D., & Ismaiel, H. S. (2005). Comparing and unifying slope estimates across psychometric function models. *Perception & Psychophysics, 67*(7), 1289-1303.
Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *The Journal of Neuroscience, 36*(36). 9420-9434.
Greene, E. & Dodge, M. (1995). The influence of prior record evidence on juror decision making. *Law and Human Behavior, 19*(1), 67-78.
Hausmann, D., Zulian, C., Battegay, E., & Zimmerli, L. (2016). Tracing the decision-making process of physicians with a Decision Process Matrix. *BMC Medical Informatics and Decision Making, 16*, 133.
Horowitz, I. A. (1997). Reasonable doubt instructions: Commonsense justice and standard of proof. *Psychology, Public Policy, and Law, 3*(2/3), 285-302.
Horowitz, I. A. & Kirkpatrick, L. C. (1996). A concept in search of a definition: The effects of reasonable doubt instructions on certainty of guilt standards and jury verdicts. *Law and Human Behavior, 20*(6), 655-670.
In re Winship. (1979). 397 U.S. 358.
Izzett, R. R. & Leginski, W. (1974). Group discussion and the influence of defendant characteristics in a simulated jury setting. *The Journal of Social Psychology, 93,* 271-279.
Johnson, R. D. (1987). Making judgements when information is missing: Inferences, biases, and framing effects. *Acta Psychologica, 66*(1), 69-82.
Kagehiro, D. K. & Stanton, W. C. (1985). Legal vs. quantified definitions of standards of proof. *Law and Human Behavior, 9*(2), 159-178.
Kahneman, D. & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.
Kaplan, M. F. & Kemmerick, G. D. (1974). Juror judgment as information integration: Combining evidential and non evidential information. *Journal of Personality and Social Psychology, 30*(4), 493-499.
Kassin, S. M. & Sommers, S. R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *PSPB, 23*(10), 1046-1054.
Kerr, N. L. (1978). Severity of prescribed penalty and mock jurors' verdicts. *Journal of*

Personality and Social Psychology, 36(12), 1431-1442.
Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception and Psychophysics, 63*(8), 1421-1455.
Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131*(2), 241-254.
Kyckelhahn, T. (2015). Justice Expenditure and Employment Extracts, 2012-Preliminary. U.S. *Bureau of Justice Statistics*. NCJ 248628. www.bjs.gov
Landy, D. & Aronson, E. (1969). The influence of the character of the criminal and his victim on the decisions of simulated jurors. *Journal of Experimental Social Psychology, 5,* 141-152.
Laudan, L. (2003). Is reasonable doubt reasonable? *Legal Theory, 9,* 295-331.
Lederman, L. (1999). Which cases go to trial: An empirical study of predictors of failure to settle. *Case Western Reserve Law Review, 49*(2), 315.
Lieberman, J. & Arndt, J. (2000). Understanding the limits of limiting instructions: Social psychological explanations for the failures of instruction to disregard pretrial publicity and other inadmissible evidence. *Psychology, Public Policy, and Law, 6*(3), 677-711.
Linares, D. & Lopez-Moliner, J. (2017). quickpsy: An R package to fit psychometric functions for multiple groups. *The R Journal, 8*(1), 122-131.
McCauliff, C. M. A. (1982). Burdens of proof: Degrees of belief, quanta of evidence, or constitutional guarantees? *Vanderbilt University Law Review, 35*. 1293.
McGovern, D. P., Roach, N. W., & Webb, B. S. (2014). Characterizing the effect of multidirectional motion adaption. *Journal of Vision, 14*(13), 1-16.
Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS Computational Biology, 11*(6).
Newman, J. O. (1993). Beyond "Reasonable Doubt". *New York University Law Review, 68*(5), 979.
Park, K., Seong, Y., Kim, M., & Kim, J. (2016). Juror adjustments to the reasonable doubt standard of proof. *Psychology, Crime, and Law, 22*(6), 599-618.
Pennington, N. & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51(2). 242-258.
Pickel, K. (1995). Inducing jurors to disregard inadmissible evidence: A legal explanation does not help. *Law and Human Behavior, 19*(4), 407.
Pilly, P. K. & Seitz, A. R. (2009). What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision Research, 49,* 1599-1612.
Plasencia, C., Alderman, B., & Baron, A. (1992). A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. *Medical Decision Making.*
Shen, F. X., Hoffman, M. B., Jones, O. D., Greene, J. D., & Marois, R. (2011). Sorting guilty minds. *New York University Law Review, 86*. 1306.
Simon, R. J. (1966). Murder, juries, and the press. Does sensational reporting lead to verdicts of guilty? *Transaction, 3,* 40-42.
Simon, R. J. & Mahan, L. (1971). Quantifying the burdens of proof: A view from the bench, the jury, and the classroom. *Law and Society Review, 5*(3). 319-330.
Spackman, M. P., Belcher, J. C., Cramer, L., & Delton, Y. (2006). A qualitative investigation of

mock-jurors' theories of emotion and reason. *Cognition and Emotion, 20*(5). 671-693.
Spieser, L., Servant, M., Hasbroucq, T., & Burle, B. (2017). Beyond decision! Motor contribution to speed-accuracy trade-off in decision-making. *Psychonomic Bulletin & Review, 24*(3), 950-956.
Sue, S., Smith, R. E., & Caldwell, C. (1973). Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. *Journal of Applied Social Psychology, 3*(4), 345-353.
Tanford, S. & Cox, M. (1988). The effects of impeachment evidence and limiting instructions on individual and group decision making. *Law and Human Behavior, 12*(4), 477-497.
Thompson, W. C., Fong, G. T., & Rosenhan, D. L. (1981). Inadmissible evidence and juror verdicts. *Journal of Personality and Social Psychology, 40*(3), 453-463.
Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. I., Ginther, M. R., Jones, O. D., & Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience, 17*(9). 1270-1278.
Towers Watson. (2012). 2011 Update on U.S. Tort Cost Trends. *Willis Towers Watson.* Towerswatson.com
Van Wezel, R. J. A. &, K. H. (2002). Motion adaptation in area MT. *Journal of Neurophysiology, 88,* 3469-3476.
Vidmar, N. (1972). Effects of decision alternatives on the verdicts and social perceptions of simulated jurors. *Journal of Personality and Social Psychology, 22*(2), 211-218.
Wagenmakers et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57.
Wagner, P. & Rabuy, B. (2017). Following the money of mass incarceration. *Prison Policy Initiative.* Prisonpolicy.org
Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond "p<0.05". *The American Statistician, 73.*
Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology, 62*(5), 739-752.
Wichmann, F. A. & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics, 63*(8). 1293.
Wichmann, F. A. & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception and Psychophysics, 63*(8), 1314.
Wissler, R. L. & Saks, M. J. (1985). On the inefficacy of limiting instructions. When jurors use prior conviction evidence to decide on guilt. *Law and Human Behavior, 9*(1), 37.
Wistrich, A., Guthrie, C., & Rachlinski, J. (2005). Can judges ignore inadmissible information? The difficulty of deliberately disregarding. *153 University of Pennsylvania Law Review,* 1251.
Woolever, D. (2008). The art and science of clinical decision making. *Family Practice Management, 15*(5), 31-36.
Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of belief in moral judgments. *PNAS, 107*(15). 6753-6758.