DEBIASING THE COURTROOM:
USING BEHAVIORAL INSIGHTS TO AVOID AND MITIGATE COGNITIVE BIASES

by

David V. Yokum

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PSYCHOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2014
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by David V. Yokum, titled “Debiasing the Courtroom: Using Behavioral Insights to Avoid and Mitigate Cognitive Biases,” and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Christopher T. Robertson  
Date: 11/17/2014

Shaun Nichols  
Date: 11/17/2014

Massimo Piattelli-Palmarini  
Date: 11/17/2014

Terry Connolly  
Date: 11/17/2014

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: Lynn Nadel
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: David V. Yokum
“None of us got where we are solely by pulling ourselves up by our bootstraps. We got here because somebody—a parent, a teacher, an Ivy League crony or a few nuns—bent down and helped us pick up our boots.”

- Thurgood Marshall

I have benefited from a remarkable set of teachers who have taught me ways of thinking empirically about how the world might work, pressed me to engage in a meaningful life, and inspired me by their examples of moral and happy relationships. To my elementary teacher Sister Elizabeth, for demanding precision in language. To my high school science teacher Allison Tupman, whose unending support sparked a confidence to start asking questions. To Dr. Rob Rockhold, brave enough to let a teenager into his pharmacology laboratory. To my college professor Wayne Shew, whose palpable passion for teaching irresistibly spread a joy for learning. To Lois LaCivita Nixon, who put me in front of classrooms to teach, and who showed me that art and literature and emotion can and must infuse the pursuit of science—my PowerPoints have never been the same. And to the other teachers too numerous to mention, you have sparked a curiosity and passion in my life. I am deeply grateful, and forever changed.

My mentors at the University of Arizona have been unparalleled. I first met Shaun Nichols, Massimo Piattelli-Palmarini, and Terry Connolly over lunch when interviewing for the doctoral program. It was a conversation that solidified my decision. I’ve benefited from their advice, critical feedback, and guidance ever since, not to mention an occasional laugh over beer. Lynn Nadel, my dissertation advisor at the psychology department, has been a constant advocate and source of support. His wide and deep knowledge of the field is truly inspiring, and I hope to one day develop a similar wisdom. All of you have pushed me to be a better scientist and provided exemplars for my professional development. You are all friends. I thank you sincerely.

I am especially grateful for the remarkable mentor, exceptional colleague, and dear friend, Chris Robertson. We met at a time when my passion for academic research was waning; our first collaboration, however, re-kindled the excitement of discovery. He has a natural talent for cutting to the core of an issue, for providing a fresh perspective on how to answer old problems. My thinking has been sharpened by his guidance, and my enjoyment of graduate school lifted up by his good humor. Without doubt, the trajectory of my career has been forever enriched. I look forward to our collaborations ahead. Thank you for everything, Chris.
A special thanks to Beth Owens, without whom I would never have made it through the administrative weeds. Your patience and good humor have saved me many times.

To my classmates, students, and friends, you have made the journey fun. To Trevor Kvaran, for jokes over lunch sushi and the ability to have clever conversations about basically any topic. To Brandon Terrizzi, for inspiring me with his genuine passion for learning, and for making me laugh endlessly. To Eddie Sue, for action on the soccer pitch. To Filippo Rossi, for beer at 1702, endless ping-pong matches, fun debate, and being a fantastic colleague. Grazie, il mio caro amico.

Michael and Jeff, my brothers, have always been mentors in my life. The occasional wedgie aside, you’ve both played critical parts in shaping who I am, and I admire the men you’ve become. I am extremely fortunate to have you and your families in my life.

To the animals in my life, Keylee, Chuey, Mouse, Godzilla, Philo, Teito, and Mr. Bubbles. And most of all to the zookeeper, Sara, my wonderful wife. At times in writing this work, your comfort and high spirits made all the difference.

My father—my dad—Edward Yokum, passed away while I was preparing this dissertation. He had an insatiable appetite for reading, tinkering with facts and objects, and thinking deeply about the purpose of life. He was always eager to hear what I was learning, to press me for more, or to crack a joke. I like to think that some of your thoughtfulness and passion for living the good life rubbed off on me. I think often of you, and remain inspired to ensure my life be one that would make you proud.

My mom, Judi Yokum, was also—and still is—always book in hand (or rather Kindle in hand). Her support and belief in me has been unwavering, a critical component of my success. Together my parents have shown that knowledge is only a part of the puzzle; more important is what you do with it. How do you use your good fortune to make the world better? I have a lot more living to do trying to figure that out, but I am fortunate to have parents that instilled in me the importance of trying to do so. Mom and Dad, you have shaped my life, in endless ways, and to the extent I show any wisdom, it no doubt traces back to your influence. Thank you, with love.

David Yokum
Tucson, Arizona
November 11, 2014
DEDICATION

“Gratitude is one of the least articulate of the emotions, especially when it is deep.”

– Felix Frankfurter

For my wife, Sara, a remarkable woman—and amazing teacher herself—without whom I could never have completed this journey, nor would it have been nearly as fun and exciting. Thank you for everything; words cannot thank you enough. I look forward to the next decade and beyond, and I promise this is the last degree.
TABLE OF CONTENTS

ABSTRACT ................................................................................................................ 13

CHAPTER 1: INTRODUCTION .................................................................................... 14

CHAPTER 2: THE NONSENSE EXPERT EFFECT ............................................................. 18

Abstract ................................................................................................................... 19
Background .............................................................................................................. 20
Method .................................................................................................................... 22
Participants .................................................................................................... 23
Materials and procedure ............................................................................... 24
Results ..................................................................................................................... 29
Verdict ............................................................................................................ 29
Perceptions of the witness ............................................................................ 32
Discussion ................................................................................................................ 34
Conclusions .............................................................................................................. 41
References ............................................................................................................... 41
Attachment A. Trial Stimulus .................................................................................. 46

CHAPTER 3: THE BIAS BLIND SPOT DURING VOIR DIRE ............................................. 56

Abstract ................................................................................................................... 57
Background .............................................................................................................. 59
The fair trial guarantee .................................................................................. 60
The psychology behind self-diagnosing bias ................................................. 62
The Skilling test for bias: Juror self-diagnosis ................................................ 66
General Method ...................................................................................................... 71
Analytic strategy ............................................................................................ 71
The biasing stimuli: Pretrial publicity (PTP) ................................................... 73
Part I: The inability to self-diagnose bias due to PTP .............................................. 75
Experiment 1: Law student self-diagnosis ..................................................... 75
Method ........................................................................................................... 75
Participants ......................................................................................... 75
Procedure ............................................................................................... 76
Biasing stimuli ........................................................................................ 76
Operationalizing the Supreme Court’s voir dire protocol ....................... 77
Trial stimulus ......................................................................................... 79
Outcome measures ................................................................. 80
Results .............................................................................................. 80
The effect of pretrial publicity ...................................................... 80
The accuracy of jurors’ self-diagnoses ........................................ 82

Experiment 2: National sample self-diagnosis ................................ 85
Method .............................................................................................. 85
Participants .................................................................................. 85
Procedure ................................................................................... 87
Results .............................................................................................. 88
The effect of pretrial publicity ...................................................... 88
The accuracy of jurors’ self-diagnoses ........................................ 89

Discussion ............................................................................................. 91
Does juror self-diagnosis help at all? ........................................... 91
Opinion-screening as an alternative to the Skilling protocol ....... 92
Unexplained behavior of fair jurors disqualifying themselves ... 95

Part I: General discussion ................................................................. 97

Part II: Self versus other diagnoses ................................................ 101
The bias blind spot ...................................................................... 102

Experiment 3: Self vs. other (undergraduates) ............................... 104
Methods ............................................................................................. 105
Participants .................................................................................. 105
Procedure ................................................................................... 106
Results .............................................................................................. 107
The effect of pretrial publicity ...................................................... 107
First versus third person diagnoses .......................................... 108
First versus third person diagnostic accuracy ......................... 109
Discussion ............................................................................................. 110

Experiment 4: Self vs. other (MTurk) ................................................ 111
Method ............................................................................................. 111
Participants .................................................................................. 111
Procedure ................................................................................... 112
Results .............................................................................................. 112
The effect of pretrial publicity ...................................................... 112
CHAPTER 7: THE NOVEL NEW JERSEY EYEWITNESS INSTRUCTION INDUCES SKEPTICISM BUT NOT SENSITIVITY ................................................................. 268

Abstract ................................................................................................................. 269
Background............................................................................................................. 270
A uniquely new instruction........................................................................ 272
Method.................................................................................................................. 275
Design and stimuli........................................................................................ 275
Participants .................................................................................................. 280
Procedure and measured variables.................................................................. 281
Results ................................................................................................................... 282
Verdict.......................................................................................................... 282
Self-Perceived comprehension, confidence, and influence........................ 286
Discussion.............................................................................................................. 287
Attachment E. New Jersey (“enhanced”) jury instructions......................... 295
Attachment F. Florida-based (“standard”) jury instructions ......................... 301
Attachment G. Additional jury instructions .................................................... 302
References............................................................................................................. 303

CHAPTER 8: CLOSING .............................................................................................. 306

The emergence of applied behavioral science................................................. 307
The White House Social and Behavioral Science Team (SBST) .................. 309
Example 1: Industrial funding fee.................................................................. 310
Objective and background ............................................................................. 310
Research insight ............................................................................................... 310
Methodology ........................................................................................................ 312
Results ................................................................................................................ 312
Next steps .......................................................................................................... 312
Example 2: Influenza vaccine outreach........................................................... 313
Objective and background ............................................................................. 313
Research insight ............................................................................................... 313
Methodology ........................................................................................................ 314
Results ............................................................................................................. 318

Example 3: Curbing CMS fraud, waste, and abuse ........................................... 318

Objective and background ............................................................................. 318

Research Insight ........................................................................................... 319

Methodology ................................................................................................... 319

Results ............................................................................................................. 321

Next steps ....................................................................................................... 321

Applied behavioral science on the horizon .................................................... 321

References ...................................................................................................... 323

REFERENCES .................................................................................................... 326
ABSTRACT

How can empirical science, and psychology in particular, be harnessed to avoid or eliminate unwanted biases? The body of work herein explores this question across twelve experiments. The first approach we consider is placing the onus on the individual to root out any already existing bias within him or herself. Chapter 3, for example, presents experiments that assess whether people (viz., jurors during voir dire) can accurately “self-diagnose” when they are irreparably biased by negative pretrial publicity. (The answer is a resounding no). A second approach is to try and avoid letting bias enter the courtroom in the first place. Chapter 4, for example, provides an experimental test of an institutional solution known as blind expertise, wherein certain biases of an expert witness are avoided by having an intermediary pick the expert, and then having the expert render an opinion before knowing which litigant made the request. In Chapter 7, we consider a third approach to handling bias, one that concedes it will exist in the courtroom. Namely, instruct jurors on the existence of bias, so that they can try to weigh it properly. To this end we test a recently enacted New Jersey instruction on eyewitness testimony. We find that jurors do not become more sensitive to low versus high evidence quality, but instead they discount the eyewitness testimony across the board. Across this inquiry, we deploy several novel tactics; in Chapter 5, for instance, we explore how continuous response measurement (CRM) can provide unique insights into the study of reasoning, and in particular how jurors parse trial evidence. We end in chapter 8 with a more general discussion of how behavioral science can be applied across law and policy.
CHAPTER 1: INTRODUCTION
The body of work herein explores the use of behavioral science within the legal system. Many of the theoretical findings and developed methods are, of course, much more broadly applicable. But as a frame of reference, we will explore this general question: How can empirical science, and psychology in particular, be harnessed to eliminate or avoid unwanted biases in during legal proceedings?

The second chapter, on the “nonsense expert effect,” is a simple demonstration of a courtroom bias. We find that jurors are unduly swayed by the presence of incomprehensible, unnecessary mathematical formulae. There are many other biases that one might choose to demonstrate within the courtroom, but this one suffices to motivate the inquiry.

What can be done about courtroom bias? The first approach we consider is placing the onus on the individual to root out any already existing bias within him or herself. The courts place this burden on people at several turns: it asks jurors during voir dire whether they can be fair and impartial, for example, and it asks the same of judges with regards to judicial disqualification. Chapter 3 presents four experiments that assess whether people (viz., jurors during voir dire) can accurately “self-diagnose” when they are irreparably biased, in this instance by negative pretrial publicity. (The answer is a resounding no). A solution based on the idea of a bias blind spot is proposed—namely that people might do better assessing whether the same information they were exposed to would bias others—and shows promise. Regardless the key lesson is that self-diagnoses are unreliable.
A second approach is to try and avoid letting bias enter the courtroom in the first place. For example, blind expertise is an institutional solution that aims to simply avoid many of the financial and social pressures that can bias expert witnesses. The idea is to have an intermediary pick the expert, and then have the expert render an opinion before knowing which litigant made the request. But for this benefit to be felt in the courtroom, jurors must actually perceive the blinded expert to be more persuasive than the traditionally hand-picked and coached expert. Chapter 4 provides an experimental test of precisely this issue, and finds that blinded experts are highly persuasive witnesses.

We take a methodological detour in Chapter 5, to explore how continuous response measurement (CRM) can provide unique insights into the study of reasoning, and in particular how jurors parse trial evidence. The dataset and opportunity for learning are rich, but work needs to be done establishing rigorous methods of data collection and analyses. Chapter 5 creates some of the framework, and begins developing methods for handling the data.

Chapter 6 replicates the blinded expert experiment in a diverse community sample, and also couples it with CRM. We see once again the substantial benefits of using a blinded expert, and also begin to unpack—with the aid of CRM—how it renders enhanced credibility via both a halo effect and by how jurors weigh an expert’s later testimony. We see other benefits of CRM as a method as well; for example, we uncover a particularly powerful moment of the trial that jurors, due to either inability or
unwillingness to report, never mentioned when asked to provide after-the-fact rationales for their decision.

In Chapter 7, we consider a third approach to handling bias, one that concedes it will exist in the courtroom. Namely, instruct jurors on the existence of bias, so that they can try to weigh it properly. To this end we test a recently enacted New Jersey instruction on eyewitness testimony, one that aims to empower jurors to discern reliable from erroneous eyewitness testimony. We find that jurors do not become more sensitive to low versus high evidence quality, but instead they discount the eyewitness testimony across the board.

All told, we will explore nine different experiments, with close to 2,000 participants drawn from a variety of in-person and online populations, spanning a wide variety of theoretical and applied concepts. My hope is the reader will emerge with a palpable sense of how much experimental psychological has to contribute to law and policy. I return to this possibility in the final chapter, Chapter 8, to briefly mention a new endeavor (and three additional experimental examples) with the White House, where we are beginning to apply behavioral insights to law and policy on a national scale.
CHAPTER 2: THE NONSENSE EXPERT EFFECT

Author Note

Thanks to Corinne Carson, Jessica Gerson, and Rachel Dyckman for assistance implementing the experiment, and to Christopher Robertson for conceptual guidance.
Abstract

Erisksson recently demonstrated a “nonsense math effect,” wherein readers judged academic abstracts to reflect higher quality research if a mathematical statement was included—even though, notably, the included model ($T_p = T_0 - fT_0d_j^2 - fT_p d_j$) was completely meaningless (see 2012, 7(6) JDM 746). We test here for a “nonsense expert effect” more generally, and try to unpack whether audiences are allured by other types of expert complexity, or whether there is something special about math. (We also fix a possible confound from the Eriksson study). Subjects ($N = 317$) were cast as mock jurors in a patent case. We manipulated two forms of nonsense complexity—linguistic (erudite or simple) and mathematical (none or a convoluted math equation) — using a 2×2 between-subjects design. Verdict rates were nearly identical regardless of whether simple or erudite language was used: 72% versus 71%, respectively, a difference of only -1%, 95% CI [-.11, .09], $p = .87$. The presence of convoluted math equations, however, had a substantial effect. The plaintiff win rate increased from 65% with no equations to 77% with equations—an absolute shift of 12%, 95% CI [.02, .22], $p = .02$.

Key words: jury research, judgment and decision making, nonsense math effect, expert witness
Background

Are audiences unduly persuaded or dissuaded by experts who dress their commentary up with needlessly erudite jargon or unnecessary use of convoluted equations? We explore this question in the applied context of the courtroom, although the answer has many other implications. It is an issue that fits within a larger inquiry, of which there has been research, concerning the ability of jurors to parse through complicated information within the courtroom. Researchers and policy-makers have, for example, grappled with jurors’ ability (or lack thereof) to comprehend legal concepts and judicial instructions (Lieberman & Sales, 2000). They have scrutinized whether laypeople properly understand the technical concepts and methodologies behind scientific testimony (Cooper, Bennett, & Sukel, 1996). And they have probed how particular types of scientific evidence, notably polygraph tests, blood typing, DNA matching, or other forensic methods of the “CSI” variety, are weighed by jurors (Devine, 2012, pp. 136–145). Yet our inquiry is not exactly whether people comprehend the evidence, and what they do to cope if they cannot. We instead ask a more direct question about the complexity itself: does an expert, by the mere fact of couching his or her commentary in complexity, gain extra persuasion oomph?

The inspiration for our particular question comes from a short experimental paper by Kimmo Eriksson on the “nonsense math effect” (2012). Participants, all of whom had postgraduate degrees, were asked to rate the quality of two scientific abstracts, one about food-sharing practices among a foraging tribe and another on the employment consequences of incarceration. The manipulation was whether or not a
completely unrelated mathematical statement, namely, that “[a] mathematical model
\(T_{PP} = T_0 - f T_0 d_f^2 - f T_p d_f\) is developed to describe sequential effects,” was jammed
at the end of the abstract. It was total gibberish: the abstracts had nothing to do with
sequential effects, and none of the equation symbols had any reasonable
correspondence to the text. Nonetheless, most readers rated the abstracts with such
nonsense math as of higher quality (about 5 points higher on a 100 point scale).

The nonsense math effect seems to cut against the grain of almost all other
advice about how to be persuasive. Style manuals, for example, overwhelming press for
simplicity. Strunk and White (1999) considers it an “elementary principle” that one
should “omit needless words,” and the APA Publication Manual advises that authors
should “say only what needs to be said” and “eliminat[e] redundancy, wordiness,
jargon, evasiveness, overuse of the passive voice, circumlocution, and clumsy prose.”

Harry Reasoner (a prominent litigator with multiple $100 million-plus jury awards under
his belt), stresses that “[t]he key to winning is being able to simplify in a clear and
powerful way. It’s the single most important thing to accomplish at trial” (1992, p. S8).
And at least according to self-reports, jurors seem to believe that experts are more
compelling if they can convey technical information non-technically (e.g., Champagne,
Shuman, & Whitaker, 1992; Shuman, Whitaker, & Champagne, 1994).

One possibility is that, in fact, there is no nonsense math effect. Arguably the
manipulation in Eriksson—namely, the extra sentence that “[a] mathematical model
\(T_{PP} = T_0 - f T_0 d_f^2 - f T_p d_f\) is developed to describe sequential effects”—is more
than just math. The talk about developing a model suggests extra work, and thus an
abstract reader might reasonably conclude that there is more substance behind the
project. After all, there is much substance that cannot be fully explained in a word-
limited abstract, so the fact a reader would give the benefit of doubt is not exactly
unreasonable. The reader might be legitimately curious to learn more about these
sequential effects at the conference. But the situation is different in, say, a courtroom.
There the full scope of facts must come out; litigants do not refer to models they will
show jurors at some other time and place.

It remains to be seen, then, whether something like a nonsense math effect will
emerge in an experimental design that more cleanly isolates the math and puts it in an
applied setting where, unlike the abstract reader, the audience is presented the full
facts. We test that proposition here, in a civil patent case. We also aim to further
extend the inquiry by testing, not only math complexity, but also linguistic complexity.
Might a similar dynamic emerge simply because an expert uses technical jargon?

**Method**

A 2×2 between-subjects factorial design was fielded. Participants read one of
four versions of an abbreviated trial transcript of a patent case. The substantive content
was identical across vignette versions, but the manner in which the plaintiff’s expert
witness delivered that content in testimony was varied along two dimensions: language
(simple or erudite) and use of mathematical equations (none or convoluted). Individual
verdict preference was the primary outcome measure.
2. Nonsense expert effect

Participants

Jury eligible adults were recruited via Amazon Mechanical Turk, an online labor market, in the spring of 2013. (See Paolacci & Chandler, 2014, for a review of the “MTurk” platform and its common use in research). They were offered $0.75 in exchange for completing an approximately 15-20 minute online task. Of 364 persons submitting a completed questionnaire, 47 (or 13%) were removed for wrongly answering one or more of three blatant attention check questions, such as: “For quality assurances purposes, please select ‘somewhat agree.’” (See Berinsky, Margolis, & Sances, 2014, regarding use of such “screener” questions; see also Paolacci, Chandler, & Ipeirotis, 2010, p. 416, finding that the rate of screener failure was indistinguishable across MTurk and college samples).

The final sample ($N = 317$) was, relative to 2010 U.S. Census data, disproportionately white (82%) and male (57%), as well as younger ($Mdn = 29$, $M = 33$, $SD = 11$) and more educated (38% had a college degree and an additional 10% had college plus graduate degrees). Chi-square and $t$-test checks did not reveal any notable differences in these demographic variables across experimental groups, and regardless there is no reason to expect (nor evidence in the data here) that race, sex, or age per se modifies any effects of testimonial complexity as manipulated in this study. What might be relevant, however, is a person’s educational obtainment or enjoyment of cognitive effort, either of which could affect how a person responds to complex scientific testimony. To this end we will further examine below the education variable and scores from a scale known as Need for Cognition.
Materials and procedure

Participants completed the experiment via the online survey platform Qualtrics (see http://www.qualtrics.com/). The task opened with an informed consent page and demographics questionnaire, followed by task instructions and an admonition to “[p]lease read carefully and approach this case with the same seriousness and diligence that you would use in a real trial.” The next click directed the participant into the start of the trial vignette.

The core of the vignette, read by all participants, was based on the 2007 Federal Circuit Court of Appeals case *Paice LLC v. Toyota Motor Company* (2007). The names in the vignette were altered to *Johnson v. SMT* to minimize recognition of the case or parties (either from memory or internet search). At issue was an alleged infringement of the plaintiff’s patent, which detailed a high-voltage method of powering gas-electric hybrid vehicles. The patent specified a drivetrain arrangement—one using a microprocessor and controllable torque transfer unit (CTTU)—capable of accepting rotational force (torque) from both an internal combustion engine and an electric motor in a controlled fashion. It is drivetrain components that deliver power to driving wheels, and the value behind the technology was figuring out how to effectively coordinate and deliver power from two different engines, one combustive and the other electric. The dispute boiled down to whether the defendant’s drivetrain was substantially equivalent to the plaintiff’s patented drivetrain.

Attachment A provides the full survey stimulus. Across all versions, the case opened with instructions from Judge Roberts, who highlighted the purpose of the trial
and provided jury instructions, with examples, on the legal meaning of patent infringement and what the plaintiff must prove in order to prevail on such a claim. A question intervened to assess participant comprehension of those instructions. The vignette then provided additional case context, such as that the patent in dispute was issued in 1994, SMT’s “Planetary Gear” device was alleged to have infringed that patent, and that both expert witnesses who would speak to the matter were “qualified by education and experience as experts in both patent law and mechanical engineering.” Testimony from the plaintiff’s and then the defendant’s expert witnesses came next. The vignette ended with closing jury instructions from Judge Roberts, who explained (among other things) that the plaintiff faced a preponderance of evidence standard, meaning the plaintiff must prove it was “more likely true than not true” that the defendant infringed the claim at issue.

The experimental manipulations were embedded in the plaintiff’s expert witness testimony. All other aspects of the trial transcript were verbatim across the four versions. The first independent variable was language complexity: the witness used either a “simple” or “erudite” speaking style. Consider the following, in response to the question, “Dr. Matthews, could you please explain to the Court the patent at issue in this case?” as examples:

**Simple Language:**

“Johnson’s patent relates to the drivetrain in hybrid vehicles. The drivetrain in an automobile is very important as it sends power from the engine to the wheels, and controls the amount of torque. To break it down even further, the torque in an automobile relates to a force rotating an object around an axis, and can be thought of as a twist to an object, so when there is more torque, there is more power driving the wheels of the car. Now, in a “typical” gas automobile,
the wheels are driven using torque that is supplied only by an engine. However, in hybrid vehicles, the wheels are driven using torque that is supplied by an engine, an electric motor, or a combination of the two.”

Complex Language:

“Johnson’s patent relates to drivetrains for hybrid electric vehicles. In a conventional fuel-powered automobile, the wheels are driven using torque supplied only by an internal combustion engine, or “ICE”. In hybrid electric vehicles, the wheels are driven by means of torque supplied by an ICE, an electric motor, or an amalgamation of the two.”

As can be seen, linguistic complexity is roughly operationalized in terms of vocabulary and sentence structure. The simple witness uses words like “combination,” while the erudite witness says “amalgamation”; and for the unavoidable technical word “torque,” the simple witness explains that “torque in an automobile relates to a force rotating an object around an axis, and can be thought of as a twist to an object,” while the erudite witness does not. An objective if imprecise measure of the readability of the two versions can be obtained from the Flesch-Kincaid index. This is a weighted measure of average sentence length and the number of syllables per word, which is then mapped onto a grade level in the United States education system. The Flesch-Kincaid scores were 12.8 for the simple version and 14.5 for the erudite version.

The second independent variable was math complexity, operationalized by either the presence or absence of a convoluted set of equations (coupled with a small amount of textual explanation) defining, and then redefining in a circular manner, the formula for torque. In particular, the plaintiff’s expert merely shows the mathematical definition of torque, $\tau = rF \sin \theta$, and then notes that it can also be redefined in terms of the patented device as $\tau_{wheels} = F_{TTU}(\tau_{motor}\tau_{MG1})$. He drones on that this function
can also be rewritten as $\tau_{wheels} = a_1 \cdot \tau_{motor} + a_2 \cdot \tau_{MG1}$. Another seven equivalent equations are then rattled through. Note, importantly, that these equations, although technically accurate, provide no information above and beyond what the plaintiff’s expert already said verbally. This is particularly true to the extent that the equations are all equivalent algebraic expressions of the same underlying formula.

After one of the four plaintiff’s expert witness testimonies—the cross of language (simple or erudite) and math (none or convoluted)—the script converged again for all participants. There was an additional comment by the plaintiff’s expert witness asserting his opinion this was “a clear case of infringement,” then testimony from the defendant’s expert witness, and finally legal instructions read by Judge Roberts.

The verdict form was presented next. It asked: “Has the Plaintiff (Johnson) proven, by the greater weight of the evidence, that the Defendant (SMT) infringed the patent?” Participants answered either “Yes, I find in favor of the Plaintiff” or “No, I find in favor of the Defendant.”

A set of questions related to the verdict decision were presented on the subsequent page. Participants were asked to write the rationale for their decision, and then indicate on a 6-point Likert scale the extent to which they strongly agreed, somewhat agreed, neither agreed nor disagreed, somewhat disagreed, or strongly disagreed with the following statements:

- The plaintiff’s expert seemed trustworthy.
- The plaintiff’s expert was persuasive.
2. Nonsense expert effect

- The plaintiff’s expert was qualified. I understood the plaintiff’s expert testimony.

- The plaintiff’s expert testimony was complex.

Finally, the Need for Cognition (NFC) short form questionnaire (Cacioppo, Petty, & Chuan Feng Kao, 1984) was administered. NFC captures the “tendency to engage in and enjoy thinking” (Cacioppo & Petty, 1982, p. 116); or as Cohen, Stotland, & Wolfe earlier put it, a “need to structure relevant situations in meaningful, integrated ways” (1955, p. 291). This personality disposition is related to the extent to which a person engages in systematic processing versus peripheral processing, that is, effortful scrutiny of the coherence and plausibility of arguments versus the use of mental shortcuts (or heuristics) to evaluate arguments (Petty & Cacioppo, 1986; see also Kahneman, 2003, describing this in terms of System I versus System II). As reviewed in Devine (2012, pp. 108–110), a variety of mock juror studies have found that high-NFC people are more sensitive to logical flaws in a case and, notable for present purposes, “better able to appreciate variation in the quality of scientific research presented by expert witnesses.” The implication here is that high-NFC participants should be more likely to scrutinize the evidence than low-NFC participants and, as a result, be more aware of the nonsense nature of the superficially erudite language or convoluted equations—neither of which provide additional information— and thus less likely to be allured into using it as a heuristic for meaningful evidence. From a theoretical perspective, an interaction effect between NFC and complexity (of either the language or math variety or both) would support the notion that any complex testimony effects emerge via a peripheral
2. Nonsense expert effect

processing pathway—in other words, jurors use complexity as a heuristic for compelling scientific testimony.

Results

Verdict

Figure 1 charts point estimates, plus confidence intervals, for the proportion of participants finding in favor of the plaintiff by experimental condition. Verdict rates were nearly identical regardless of whether simple or erudite language was used: 72% versus 71%, respectively, a difference of only -1%, 95% CI [-.11, .09], \( p = .87 \). The presence of convoluted math equations, however, had a substantial effect. The plaintiff win rate increased from 65% with no equations to 77% with equations—an absolute shift of 12%, 95% CI [.02, .22], \( p = .02 \). One way to consider the size of this effect is to note that 8% is equivalent to about one person on a 12 person jury.
2. Nonsense expert effect

Figure 1. Proportions [95% CI] of verdict preferences in favor of the plaintiff. There was no main effect of language (a difference of only -1%, 95% CI [-.11, .09], \( p = .87 \)), but math had a substantial effect (a 12%, 95% CI [.02, .22], \( p = .02 \)).

Table 1 presents results from logistic regression models predicting verdict (0 = Defense; 1 = Plaintiff) as a function of Language (Simple = 0; Erudite = 1), Math (None = 0; Convoluted = 1), the interaction of Language and Math, Need for Cognition score (split at the median, for ease of interpretation, with Low = 0 and High = 1), and highest level of educational obtainment (also simplified, with 0 = less than college and 1 = college degree or higher). Results are coded such that a positive coefficient \( \beta \) indicates a verdict in favor of the plaintiff is more likely.
Table 1. Logistic regression predicting a verdict in favor of the plaintiff.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (S.E.)</td>
<td>β (S.E.)</td>
<td>β (S.E.)</td>
<td>β (S.E.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.66 (.21)</td>
<td>.73 (.24)</td>
<td>1.56 (.51)</td>
<td>0.63 (.31)</td>
</tr>
<tr>
<td>Language_Erudite</td>
<td>-0.04 (.25)</td>
<td>-0.19 (.34)</td>
<td>-0.05 (.25)</td>
<td>.12 (.37)</td>
</tr>
<tr>
<td>Math_Convoluted</td>
<td>0.58 (.25) *</td>
<td>.42 (.36)</td>
<td>.59 (.26) *</td>
<td>.63 (.37) ø</td>
</tr>
<tr>
<td>Language_Erudite * Math_Convoluted</td>
<td>-</td>
<td>.33 (.51)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NFC_High</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.04 (.42)</td>
</tr>
<tr>
<td>NFC_High * Language_Erudite</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-.31 (.50)</td>
</tr>
<tr>
<td>NFC_High * Math_Convoluted</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-.08 (.51)</td>
</tr>
<tr>
<td>College</td>
<td>-</td>
<td>-</td>
<td>-28 (.27)</td>
<td>-</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>-</td>
<td>-.37 (.27)</td>
<td>-</td>
</tr>
<tr>
<td>White</td>
<td>-</td>
<td>-</td>
<td>.13 (34)</td>
<td>-</td>
</tr>
<tr>
<td>Age_10</td>
<td>-</td>
<td>-</td>
<td>-.21 (10) *</td>
<td>-</td>
</tr>
</tbody>
</table>

Null deviance (df) 380 (316) 380 (316) 380 (316) 380 (316)
Residual deviance (df) 374 (313) 372 (313) 367 (309) 374 (311)

Note. Change in logit probability (standard error) given by β (S.E.). ø p < .10, * p < .05.

Model one provides a regression analytic estimate of the main effects discussed before, again revealing no evidence of a language complexity effect, but that use of convoluted math equations substantially increases the plaintiff win rate by about 14%, 95% CI [.02, .27], p = .02. Model two adds an interaction term. The effect of convoluted
math is larger if alongside erudite rather than simple language, although the estimated increase of 8% is not reliably different than zero, \( p = .51, 95\% \text{ CI } [-.17, 33] \).

Sex, age, and race were collected to describe the sample and provide additional control (beyond randomization), but were otherwise not of theoretical interest. And as seen in Model 3, inclusion of these demographics as covariates does not notably change the estimated coefficients for either Language or Math. (Side observation: with each additional decade of life, participants are about 5% less likely to find for the plaintiff).

Model 4 examines Need for Cognition (NFC). High NFC persons were no more likely to find for the plaintiff than low NFC persons (the difference is 1%, 95% CI [-.19, .22], \( p = .92 \)), nor was any such difference predicted. The prediction instead was that high NFC individuals would be less affected by either erudite language or convoluted math. This is supported by the direction of observed effects, but the estimates are not reliably different from zero. The effect of erudite language is about 8-percentage points lower for high NFC persons, 95% CI [-.32, .17], \( p = .54 \), and the effect of convoluted math is about 2-percentage points lower, 95% CI [-.27, .23], \( p = .88 \).

**Perceptions of the witness**

Perceptions of the plaintiff’s expert witness and his testimony are described in Table 2.
Table 2. Perceptions of the plaintiff and his testimony. Data show means (SDs) from a 6-point Likert scale, spanning (1) strongly disagree, somewhat disagreed, neither agreed nor disagreed, somewhat disagreed, and (6) strongly agreed. Larger numbers indicate that the plaintiff’s expert or his testimony: seemed more complex; felt easier to comprehend; seemed more persuasive; or seemed more qualified.

<table>
<thead>
<tr>
<th>Language</th>
<th>Math</th>
<th>n</th>
<th>Complex</th>
<th>Comprehend</th>
<th>Persuasive</th>
<th>Qualified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>None</td>
<td>80</td>
<td>4.0 (1.2)</td>
<td>5.0 (0.75)</td>
<td>4.8 (1.0)</td>
<td>5.0 (0.9)</td>
</tr>
<tr>
<td></td>
<td>Convoluted</td>
<td>79</td>
<td>4.9 (1.2)</td>
<td>4.3 (1.1)</td>
<td>4.9 (1.2)</td>
<td>5.2 (0.9)</td>
</tr>
<tr>
<td>Erudite</td>
<td>None</td>
<td>79</td>
<td>4.7 (1.2)</td>
<td>4.5 (0.9)</td>
<td>4.7 (1.0)</td>
<td>5.1 (0.8)</td>
</tr>
<tr>
<td></td>
<td>Convoluted</td>
<td>79</td>
<td>5.1 (1.2)</td>
<td>4.3 (1.1)</td>
<td>5.0 (1.0)</td>
<td>5.3 (0.7)</td>
</tr>
</tbody>
</table>

The complexity rating can be considered a form of manipulation check. We designed the Language and Math variables to embody different degrees of complexity, and participant self-report confirms they perceived it as such. The statement that the testimony was complex garnered much stronger support in the erudite than simple condition (means: 4.9 vs. 4.4, \( p < .01 \); note the Wilcoxon-Mann-Whitney (WMW) test—a non-parametric form of ANOVA, basically—is being used for statistical inference on this and all other Likert rating tests in this section). Testimony was likewise perceived as more complex if math was injected (5.0 vs. 4.3, \( p < .01 \)).

Perceived comprehension mirrored complexity. Participants felt they comprehended the simple testimony more than the erudite testimony (\( M = 4.7 \) vs. \( M = 4.4, \ p = .01, \ WMW \)) and, likewise, testimony without math more than testimony with convoluted math (4.8 vs. 4.3, \( p < .01 \)). Although even in the erudite condition the average rating indicated “somewhat agree[ment]” that they understood the plaintiff’s testimony.
Notably, participants did not rate the expert any differently in terms of either persuasiveness or qualification as a function of language complexity. Whether simple or erudite language, the plaintiff’s expert witness hovered in persuasiveness around 4.8 ($p = .82$) and in qualification around 5.1 ($p = .39$). But participants did rate the expert differently as a function of math: persuasiveness increased from 4.8 to 4.9 ($p = .02$) and qualification from 5.1 to 5.3 ($p = .01$), when math was injected into the testimony. The shift is barely noticeable in terms of the means, but this is a misleading consequence of the Likert scale getting squished against its upper boundary—a measurement ceiling effect.

**Discussion**

Our results confirm the existence of a nonsense expert effect, at least of the mathematical variety. In particular, introduction of a convoluted mathematical equation, one which jurors admitted merely confused them, nonetheless caused a substantial increase in plaintiff wins rates: a 12-percentage point bump, 95% CI [.02, .22], from 65% to 77%, or equivalently a 19% relative increase.

This effect is “nonsense” in two ways. The first is that jurors, by design and in line with their self-reported lack of comprehension, did not actually understand the import of the mathematical formulae, and yet nonetheless it caused a substantial impact on their verdict preferences. The second is that, even if they had understood the formulae, the equations were superfluous. The equations provided no additional information above and beyond the verbal testimony.
2. Nonsense expert effect

Note that this is related to, but ultimately different from, a different thread of research on how jurors perceive superfluous brain scan imaging (e.g., imagery from functional magnetic resonance imagining, or fMRI). In particular, an early study by Gurley and Marcus (2008) found that mock jurors were more likely to reach a verdict of not guilty by reason of insanity (NGRI) if psychiatric diagnoses were coupled with anatomical brain images (see also McCabe & Castel, 2008, showing that attaching brain images next to a newspaper article makes it appear of higher scientific quality).

These findings were likely confounded, however, by the fact that jurors seem to perceive psychiatric explanations as simply less compelling than neuroscientific explanations. It was arguably not the neuroimage per se that caused more NGRI verdicts, but rather the difference in underlying substantive testimony and, in particular, the fact that neurobiological diseases seem more “real” than psychiatric ones to many jurors. An experiment by Schweitzer and Saks (2011) aimed to more cleanly isolate the imagery effect by manipulating whether the testimony was from a psychologist or neuroscientist and, if the latter, whether it was paired with a brain scan, bar graph, or a control image of an empty courtroom. In line with the confound critics, they found that the neuroscientist’s verbal testimony was more persuasive than the psychologist’s, but that the NGRI rate was similar whether the neuroscientific testimony was paired with a brain scan, bar graph, or control image. All told the balance of evidence to date suggests that neuroimagery does not exert any undue influence above and beyond the verbal testimony—it is window dressing without any effect (see Roskies, Schweitzer, & Saks, 2013, for a critical review).
Regardless it is difficult to generalize from the neuroimage research to our inquiry here. To begin with, the early results are mixed, and juror intuitions about mind-brain dualism muddy the interpretive waters. But even if those issues are sorted out, a brain scan, whatever effect it may or may not exert, would be via different mechanisms than either erudite language or mathematical formulae. Unlike the latter, a brain scan arguably seems easier, not more difficult, to understand than the underlying verbal testimony. A picture is worth a thousand words, so the saying goes. The worry about the undue influence of neuroimagery is its deceptive simplicity, that “[b]rain images resemble photographs and laypersons view them as simple and direct photographs of brain activity,” despite the fact that “brain imagining is not photography and is neither direct nor inferentially straightforward” (Roskies et al., 2013, p. 99). If jurors were overly persuaded by neuroimages, then that might reflect an unfortunate misunderstanding of how those images were created, but it would not be “nonsense” of the complexity variety explored here.

Unlike mathematical complexity, we do not find evidence that linguistic complexity allures jurors. Verdict rates stay flat around 72%. Why might this be the case?

The study of linguistic complexity within academic writing provides perspective. For one thing, we find that intuition and fact do not mesh. People often deliberately increase vocabulary and sentence complexity in the belief that such erudite language will boost their perceived intelligence. Almost 2 in 3 Stanford undergraduates, for example, admitted to “turn[ing] to the thesaurus to choose words that are more
complex to give the impression that the content is more valid or intelligent” (Oppenheimer, 2006). Academic publications—the written currency upon which scientific experts trade—are notoriously verbose and filled with jargon (see Pinker, 2014, for a fascinating discussion of why academic writing is often so bad). The basic intuition—the bamboozlement theory, as Pinker puts it—is that scholars “spout obscure verbiage to hide the fact that they have nothing to say. They dress up the trivial and obvious with the trappings of scientific sophistication, hoping to bamboozle their audiences with highfalutin gobbledygook” (Pinker, 2014; note this is not his theory for why academic writing is so bad; see also Sperber, 2010, on the “guru effect”). A less sinister proposition is that erudite language really is a useful correlate of intelligence, and thus people learn to latch onto it for good reason. Charles Spearman (1904) observed the positive correlation between intelligence and large vocabularies over a hundred years ago, and more recent studies have similarly shown an association between use of long words in class assignments, SAT scores, and exam grades (Pennebaker & King, 1999). Whatever the merits, are people actually impressed with highfalutin gobbledygook?

Research on processing fluency suggests not, or at least not always. The theory behind processing fluency is that people judge the truth and value of information partly as a function of the subjective feeling of ease associated with trying to make sense of that information (see Oppenheimer, 2008, for a review). Daniel Oppenheimer, in the aptly named article, “Consequences of erudite vernacular utilized irrespective of necessity,” provides some empirical evidence. Graduate school admissions essays were
shown in either their original language or manipulated such that certain nouns, verbs, and adjectives were replaced with their longest entry in the Microsoft Word thesaurus—either every third word was replaced ("moderate" complexity) or every word was replaced ("high" complexity). On a 14-point scale related to likelihood of accepting the candidate, highly complex essays were rated more negatively than moderately complex essays (about 2 points less), which were in turn rated more negatively than the original essays (about 1 point less). Several follow-up experiments confirmed the negative consequences of needless linguistic complexity in different domains and types of judgments (e.g., intelligence ratings after reading sociological dissertation abstracts or philosophical essays). Data such as these suggest that erudite language can backfire, so stick with simple language. But these are not the only data.

An important caveat—one of special relevance to the courtroom—is that the signal provided by processing fluency potentially varies as a function of a person's expectation regarding how fluent language should be. We sometimes expect language to be complicated. Galak and Nelson offered the example that “a student hunched over a textbook likely has different motives than does someone curled up with a mystery novel,” noting that “[g]ood textbooks are supposed to be difficult to read, a theory might hold, so an easily read book might suggest low quality” (2011, p. 250). To test this proposition, Galak and Nelson asked undergraduates to all read the same Mark Twain short story, but they manipulated expectations and processing fluency. Expectations were set by way of instructions. Subjects were informed they were participating in a “Historical Analysis Study” (and would “indicate how [they] think it fits with the time
2. Nonsense expert effect

period it was written in”) or a “Short Story Study” (and would “tell us how much you enjoyed it”). Processing fluency was manipulated by way of font; either size 12 Times New Roman size (fluent) or size 12, 1.5 condensed Times New Roman (disfluent) was used. As predicted, subjects judged the “Short Story Study” to be of higher quality when the font was fluent, but the “Historical Analysis Study” to be of higher quality when the font was disfluent. The implication is that whether simple or erudite language is more persuasive in the courtroom might depend on whether jurors expect scientific testimony to be simple or erudite.

We provide here another data point to enter into the debate, and in particular evidence suggestive that language does not have the same persuasive allure as mathematical formulae. Perhaps that allure—or rather lack thereof—has to do with what you might call “scientific veneer.” What is a scientific veneer is difficult to precisely define. The basic notion is that people have intuitions about what is scientific and what is not scientific. White coats, test tubes, and physics equations constitute stereotypes of what is scientific, whereas firefighter jackets, coffee mugs, and cooking recipes do not.¹

¹ Note that these observable details are superficial. Whether the underlying substance is also truly scientific is a separate question, one answered by assessing the used methodology (see e.g., Ayala & Black, 1993). Consider a study by Krull and Silvera (2013). A first experiment found that laypeople judged natural science techniques (e.g., electrocardiograms, or EKGs) as more “scientific” than those from behavioral science (e.g., interviews). They exploited this fact in a second experiment to test whether the superficial technique—the veneer, if you will—could make the same underlying research endeavor intuitively seem more “scientific.” For example, the statement “Dr. Miller studies cancer” was followed by either: (natural science) “To do this research, Dr. Miller uses an EKG (electrocardiogram)” ; or (behavioral science) “To do this research, Dr. Miller uses interviews.” Across this and other statements, the same research endeavor was judged more scientific if coupled with a natural rather than behavioral science techniques.
An incomprehensible statement need not have a scientific veneer. The 1998 winner of the “Bad Writing Contest” sponsored by the journal *Philosophy and Literature*, for example, spouted the following single-sentence rigmarole:

The move from a structuralist account in which capital is understood to structure social relations in relatively homologous ways to a view of hegemony in which power relations are subject to repetition, convergence, and rearticulation brought the question of temporality into the thinking of structure, and marked a shift from a form of Althusserian theory that takes structural totalities as theoretical objects to one in which the insights into the contingent possibility of structure inaugurate a renewed conception of hegemony as bound up with the contingent sites and strategies of the rearticulation of power (Dutton, 1999, citing Guggenheim Fellowship-winning professor Judith Butler).

There is nothing noticeable scientific within that train wreck of a sentence.

The math equations here might possess a scientific veneer that the language here simply does not. And perhaps that is the special ingredient. Follow up experiments could tease apart the language variable into complex grammar and vocabulary that is more or less scientific (e.g., say influenza rather than flu, or acetaminophen rather than Tylenol). A different (although not mutually exclusive) idea is that, similar to how Galak and Nelson (2011) found that the effect of processing fluency varied based on expectations, jurors here (and people more generally, for that matter) might expect any math they encounter to be difficult, and thus interpret processing disfluency as a sign of scientific rigor. This was a patent case involving engines, so one might expect (or at least not be surprised by) engineering jargon and math. But if math was not expected, such as if this was a defamation case, then perhaps the resultant disfluency would actually cause the opposite effect.
Conclusions

This is not an indictment of the juror system. It is a yellow flag that people, in general, can be misled by scientific highfalutin gobbledygook. There is no reason to expect that judges or attorneys or anyone else would do any better here. Recall the Eriksson study, showing the nonsense math effect, was conducted exclusively on persons with graduate level scientific training. And here we see no immunity from the effect as a function of either education or Need for Cognition.

This is also, we stress emphatically, not a suggestion that lawyers should start deliberately exploiting the nonsense expert effect as a trial strategy. That would be an abdication of professional responsibility. If that is not enough, it might backfire in an adversarial system. We did not test here what happens when the opposing party calls out the shenanigans of the nonsense expert. A rigorous cross-examination might effectively disarm any effect. A future experiment should test this possibility.

What is advised, however, is a heightened awareness to the subtle effects behind complex testimony. Monolithic advice that simple is always better or that jurors are always lured by technical jargon is incomplete and potentially misleading. We need more research teasing apart the different types of complexity and the mechanisms by which those types influence reasoning. And as for trial attorneys in the courtroom, be wary of excessive jargon and take action to call out nonsense math, so as to aid jurors in their task of understanding the evidence.

References

2. Nonsense expert effect


2. Nonsense expert effect


2. Nonsense expert effect


2. Nonsense expert effect


Thank you for participating in this study. Your responses are important to us; please carefully read and consider all materials.

Instructions

You will be acting as a member of the jury in this case. On the following pages, the Judge will provide instructions and introduce the case, and then you will read an excerpt from two expert witnesses' testimony on behalf of either the plaintiff or the defendant. You will then fulfill your duty as a juror by reaching a verdict. Please read carefully and approach this case with the same seriousness and diligence as you would in a real trial.

Judge's Instructions

Ladies and Gentlemen: I am Judge David Roberts. You will be asked to decide the case of Johnson v. SMT. I will briefly explain your duties as jurors.

The evidence presented will consist of the testimony of expert witnesses. In deciding the facts of this case, you should consider what testimony to accept, and what to reject. You may accept everything a witness says, part of it, or none of it.

PATENT INFRINGEMENT—GENERALLY:

This trial concerns a patent dispute. A patent gives an inventor a set of exclusive rights over her invention. She is the only person who can make, use, or sell the invention, or who can authorize others to do so. If someone else makes, uses, or sells her patented invention, she can sue them for infringement.

CLAIM INFRINGEMENT:

A patented invention will have a number of elements, which are listed in a patent in sets called “claims.” For example, a patent covering a pencil might have only one claim containing the elements of a wooden body, a graphite center, a sharpened point, and a
rubber tip. To infringe a patent, the defendant’s invention must have each of the elements of at least one claim. So, to use the example above, a pencil with a wooden body, a graphite center, and a rubber tip would not infringe—because it lacks the sharpened point. If the defendant’s device omits even a single structure or fails to perform a step recited in a claim, then you must find that defendant has not infringed that claim.

[New Page]

Judge’s Instruction

(continued. . .)

DOCTRINE OF EQUIVALENTS:

However, it is possible to infringe without having each element in the claim precisely (or “literally”) present in the defendant’s invention. Instead, the defendant can infringe if the missing element has a substantial equivalent in his apparatus. To use the example above, if the defendant’s pencil had a wooden body, a graphite center, a rubber tip, and a dull point that nonetheless enabled it to write, that dull point could be the substantial equivalent of the sharpened one. And, thus, the defendant’s pencil could infringe the plaintiff’s patent. This type of infringement is called infringement under the “doctrine of equivalents.”

Equivalence does not require that the structure be identical. One established test to determine equivalence is whether the defendant’s technology performs substantially the same function in substantially the same way to produce substantially the same result compared to the plaintiff’s patented technology.

When you think about an equivalent, you must consider whether the equivalent of an element is present in the defendant’s apparatus. You cannot consider whether the apparatus as a whole is equivalent to the plaintiff’s invention.

[New Page]

Judge’s Instruction

(continued. . .)

INFRINGEMENT DESPITE DEFENDANT’S IMPROVEMENTS:

The defendant’s apparatus may be an improvement over the invention described in the plaintiff’s patent, or contain elements not present in the patent, and nonetheless
infringe. If the defendant’s device includes every element in the plaintiff’s patent claim, it infringes despite the improvements. For example, considering the patented pencil, a pencil including a wooden body, a graphite center, a sharpened point, a rubber tip, and a foam comfort grip infringes despite the addition of the comfort grip, because all of the claim’s elements are present. However, if the improvements transform the invention to the extent that it no longer performs substantially the same function in substantially the same manner, it does not infringe.

[New Page]

Background

In 1994, the U.S. Patent and Trademark Office issued Mr. Johnson a patent for a high-voltage method of powering gas-electric hybrid vehicles. This invention was the result of years of work and research, and was one of the earliest versions of the drivetrain that most hybrid vehicles use today.

Three years after Johnson was awarded that patent, a major motor vehicle manufacturing company, SMT, released the first generation of its new line of hybrid cars.

Soon after the new car began showing up on the market, Johnson filed this lawsuit alleging that SMT’s “Planetary Gear” device infringes claim 7 of his patent, specifically concerning his controllable torque transfer unit (“CTTU”). The CTTU accepts torque input from both the gas-powered internal combustion engine (“ICE”), and the electric motor, allowing the car to seamlessly switch back and forth between gas and electric power.

[New Page]

Background

(continued. . .)

The defendants, SMT, deny that their drivetrain technology infringes Johnson’s patent because of key differences in their Planetary Gear assembly.

You, the jury, must decide whether SMT has infringed Johnson’s patent. Because this is an abbreviated trial, you will hear testimony from only one expert witness for each side. Both witnesses are qualified by education and experience as experts in both patent law and mechanical engineering.
Background

(continued . . .)

The parties have stipulated to the following facts, which means that these items are undisputed and, for the purposes of your verdict, you must assume these facts to be true.

1. Johnson owns the patent at issue and has not granted SMT permission to make, use, or sell any invention described in the patent.

2. SMT’s Senora II, Lowlander, and Hexius RX 400h 3 are hybrid electric vehicles containing two or more drive wheels receiving torque for propelling the vehicle from an output shaft, supplied by a power unit which receives power from both an ICE and an electric motor.

Johnson has sued for infringement on claim 7, which reads:

7. A hybrid electric vehicle, comprising: two or more drive wheels receiving torque for propelling the vehicle from an output shaft, and a power unit supplying drive torque to the output shaft, and comprising:
   a) an internal combustion engine (“ICE”);
   b) an electric motor;
   c) a controllable torque transfer unit (“CTTU”), which is a multi-input device or component that is controlled to transfer variable amounts of torque from the ICE and the electric motor to propel the wheels of the vehicle;
   d) a computerized control device for controlling the CTTU, the electric motor, the ICE, and the amount of relative torque supplied by each.
<table>
<thead>
<tr>
<th>Simple Language</th>
<th>Erudite Language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plaintiff’s Expert Witness</strong></td>
<td><strong>Plaintiff’s Expert Witness</strong></td>
</tr>
<tr>
<td><strong>Johnson’s [Plaintiff’s] Attorney:</strong> “Dr. Matthews, could you please explain to the Court the patent at issue in this case?”</td>
<td><strong>Johnson’s [Plaintiff’s] Attorney:</strong> “Dr. Matthews, could you please explain to the Court the patent that is at issue in this case?”</td>
</tr>
<tr>
<td><strong>Dr. Matthews [Plaintiff’s Expert]:</strong> “Johnson’s patent relates to drivetrains in hybrid electric vehicles. In a conventional fuel-powered automobile, the wheels are driven by means of torque supplied by an ICE, an electric motor, or an amalgamation of the two. In hybrid electric vehicles, the drivetrain in Johnson’s patent employs a microprocessor and a controllable torque transfer unit, which we refer to as the CTTU for short. The CTTU has lockable gears that can control the amount of torque coming in and out of the engine and the electric motor.”</td>
<td><strong>Dr. Matthews [Plaintiff’s Expert]:</strong> “Johnson’s patent relates to drivetrains for hybrid electric vehicles. In a conventional fuel-powered automobile, the wheels are driven using torque supplied only by an internal combustion engine, or “ICE”. In hybrid electric vehicles, the wheels are driven by means of torque supplied by an ICE, an electric motor, or an amalgamation of the two. This augments the level of convolution because the relative torque contributions of the ICE and the electric motor must be pooled and controlled. The drivetrain in Johnson’s patent employs a microprocessor and a CTTU that accepts torque input from both the ICE and the electric motor. In this scheme, bevel gears are furnished with microprocessor-controlled locking devices for setting the gears’ rotational freedom—if any—relative to the housing. The CTTU’s microprocessor, by virtue of its ability to control the amount of torque provided at each shaft, is capable of controlling the relative aggregates of torque transferred from the ICE and the electric motor to the drive shaft.”</td>
</tr>
<tr>
<td>“This patent focuses on the implementation in hybrid vehicles. We are looking at how torque from an electric motor is blended with that of a normal gasoline engine. The patent at issue allows an engine and an electric motor to work back-and-forth together to make the wheels drive. Johnson’s patent utilizes a microprocessor and a controllable torque transfer unit, which we refer to as the CTTU for short. The CTTU has lockable gears that can control the amount of torque coming in and out of the engine and the electric motor.”</td>
<td></td>
</tr>
<tr>
<td><strong>Johnson’s [Plaintiff’s] Attorney:</strong> “So in your professional opinion, Dr. Matthews, in what ways were the technologies similar?”</td>
<td><strong>Johnson’s [Plaintiff’s] Attorney:</strong> “So in your professional opinion, Dr. Matthews, in what ways were the technologies similar?”</td>
</tr>
<tr>
<td><strong>Dr. Matthews [Plaintiff’s Expert]:</strong> “Well in 2003, SMT began marketing their Senora II; the drivetrain in the Senora II—which is also present in their SMT Lowlander and Hexius RX 400h—is extremely similar to Johnson’s patented drivetrain that I just explained to you. Although SMT calls their drivetrain a “transaxle unit,” it still combines torque from an engine and electric motor. The gears involved in SMT’s transaxle unit work a little differently than the lockable</td>
<td><strong>Dr. Matthews [Plaintiff’s Expert]:</strong> “Well in 2003, SMT began marketing the “Senora II.” The transaxle unit of the Senora II—which is also present in another arrangement in the SMT Lowlander and Hexius RX 400h 3—is analogous to the drivetrain described in the patent in the sense that it, too, combines torque from an ICE with torque from an electric motor, the “MG2” or “traction motor”. However, instead of combining relative torque using the patent’s lockable bevel</td>
</tr>
</tbody>
</table>
gears described in the patent, but just like the patented drivetrain, SMT’s transaxle unit uses a microprocessor and gears to control the amount of torque coming from the gas engine and the electric motor, and is thus, extremely similar to Johnson’s patent.”

“...gears, SMT’s drivetrain is designed around a “planetary gear unit” or “powersplitting device,” consisting of a central “sun” gear that meshes with several “planetary” gears, which are supported by a “planetary carrier,” which in turn mesh with a peripheral ring gear.”

“In SMT’s drivetrain, the output shaft from the ICE is connected to the planetary carrier, and, thus, to the planetary gears, whereas 100% of the torque supplied by the MG2 outputs directly to the ring gear. The SMT design also employs an additional motor/generator (“MG1”) having an output shaft connected to the sun gear. Although SMT’s gear assembly transfers fractional torque, it performs the equivalent function of the patented CTTU because the microprocessor associated with SMT’s drivetrain is able to control the raw output of torque provided by the ICE and the MG2, as well as its transference to the drive shaft.

Moreover, when we specifically look at the torque, defined as: $\tau = rF \sin \delta$, the similarities between the designs become clear.
In the patented CTTU, the output torque to the wheels is described by

\[ \tau_{\text{wheels}} = F_{\text{TTU}}(\tau_{\text{motor}}, \tau_{\text{MG1}}) \]

which describes the function of the torque transfer unit (TTU). This can be rewritten as:

\[ \tau_{\text{wheels}} = a_1 \cdot \tau_{\text{motor}} + a_2 \cdot \tau_{\text{MG1}} \]

In the case of the SMT device

\[ \tau_{\text{wheels}} = F_{\text{PGU}}(\tau_{\text{motor}}, \tau_{\text{MG1}}, \tau_{\text{MG2}}) \]

or

\[ \tau_{\text{wheels}} = b_1 \cdot \tau_{\text{motor}} + b_2 \cdot \tau_{\text{MG1}} + b_3 \cdot \tau_{\text{MG2}} \]

which describes the function of the planetary gear unit (PGU). However, since in the case of the SMT device

\[ \tau_{\text{wheels}} = \tau_{\text{MG2}} \]

so we have

\[ \tau_{\text{wheels}} = b_1 \cdot \tau_{\text{motor}} + b_2 \cdot \tau_{\text{MG1}} + b_3 \cdot \tau_{\text{wheels}} \]

or

\[ \tau_{\text{wheels}} = \frac{b_1}{(1-b_3)} \cdot \tau_{\text{motor}} + \frac{b_2}{1-b_3} \tau_{\text{MG1}} \]

then replace the constants with

\[ \tau = rF \sin \theta, \text{ inputted and then outputted; the equivalence between the designs becomes clear.} \]
2. Nonsense Expert Effect
d
gives

\[
a_1 = \frac{b_1}{1-b_3}, \quad a_2 = \frac{b_2}{1-b_3}
\]

which is identical to the equation that describes the TTU. Thus, the devices are equivalent.

---

Johnson's [Plaintiff's] Attorney: “After evaluating SMT’s design and comparing it to claim seven of the patent, did you form an opinion as to whether SMT’s hybrid vehicles infringed?

Dr. Matthews [Plaintiff's Expert]: “Yes. I concluded that the defendant infringed the plaintiff’s patent. SMT’s planetary gear assembly is a multi-input device that performs substantially the same function with substantially the same result as the CTTU. While it’s true that the planetary gear assembly operates in a different way than Johnson’s CTTU, if you look at the language in claim seven, it only requires that the device be able to transfer variable amounts of torque from the engine and the electric motor to the wheels. That’s exactly what the planetary gear system does. Likewise, SMT’s microprocessor is able to dictate the amount of torque that is transferred from each power source, and thus, is equivalent to the computerized control device described in the patent. So if you go down the list of elements in claim 7, you can see that for each and every element—the electric motor, gas engine, multi-input torque transfer device, and computerized controller—the same element, or its equivalent, is present in the SMT drivetrain. Therefore, in my opinion, this is a clear case of infringement.”
Defendant’s Expert Witness

SMT’s [Defendant’s] Attorney: “Dr. Owens, how does SMT’s hybrid device differ from the patent?”

Dr. Owens [Defendant’s Expert]: “SMT’s device differs from the patent in that the planetary gear only accepts torque input from the engine, and then only 72% of that torque is transferred to the ring gear, at which point it combines with the torque inputted directly from the electric motor, or “MG2”. Therefore, the combination of torque from the engine, with the torque from the MG2 does not occur until after the 72% is output from the planetary gear to the ring gear. Thus, there is no single “device or component” that can be characterized as “multi-input” as the patent’s language requires. Moreover, the planetary gear cannot be altered from the 72/28 split, and therefore, cannot control the transfer of variable amounts of torque. To be clear, the 72/28 ratio of relative torque is constant—not variable. It doesn’t change. Thus it is not even remotely equivalent to the CTTU as described in claim 7. These features substantially differentiate SMT’s transaxle unit from the patent at issue.”

SMT’s [Defendant’s] Attorney: “After comparing the function and design of SMT’s hybrid vehicles to claim 7 of Johnson’s patent, did you form an opinion as to whether the accused vehicles infringed?”

Dr. Owens [Defendant’s Expert]: Yes I did. There is no infringement here because, as I’ve explained, neither the Planetary Gear Assembly, nor any structure in any of SMT’s hybrid cars satisfies the patent’s description of the CTTU. Nor can infringement be found under the doctrine of equivalents, because the Planetary Gear Assembly performs its task in a substantially different way than the CTTU.
Judge’s Closing Instructions

Thank you for your attention during this trial. Now it is your duty to reach a verdict based solely on the evidence presented here today. You must determine whether Johnson has proven by a preponderance of the evidence that SMT has infringed the claim at issue. A “preponderance of the evidence” means “more likely true than not true.” If, and only if, you find that SMT’s product has each of the elements present in the claim, either literally or under the doctrine of equivalents, then you must find in favor of Johnson, the plaintiff.

Remember that equivalence does not require that the structure be identical. If you find that an element in SMT’s device performs substantially the same function in substantially the same way to produce substantially the same result as an element described in the claim, then the elements are equivalent. You must assess each element individually, and if even one element (or its equivalent) as described in the claim is absent from SMT’s device, then you must find in favor of SMT.

In deciding your verdict, you may review the patent claim and consider its meaning based only on my instructions and the trial testimony.

Patent Claim 7:

7. A hybrid electric vehicle, comprising: two or more drive wheels receiving torque for propelling the vehicle from an output shaft, and a power unit supplying drive torque to the output shaft, and comprising:
   a) an internal combustion engine (“ICE”);
   b) an electric motor;
   c) a controllable torque transfer unit (“CTTU”), which is a multi-input device or component that is controlled to transfer variable amounts of torque from the ICE and the electric motor to propel the wheels of the vehicle;
   d) a computerized control device for controlling the CTTU, the electric motor, the ICE, and the amount of relative torque supplied by each.
CHAPTER 3: THE BIAS BLIND SPOT DURING VOIR DIRE

Author Note

This chapter is currently being developed for submission in collaboration with Christopher Robertson and Matt Palmer. The authors thank Joseph O'Brian, Josh Thompson, and Emily Wieser for their assistance in administering the survey in experiment three.
Abstract

Litigants are guaranteed the right to an impartial jury, one that bases its judgment only on the evidence presented in the courtroom. The Supreme Court, as recently as *Skilling v. U.S.* (2010), has instructed courts on how to screen for potentially impartial jurors: simply ask them. The empirical presumption is that jurors can self-diagnosis with sufficient accuracy. Across two initial experiments, we exposed 248 mock jurors to news articles that were either prejudicial to the defendant (in the treatment condition) or irrelevant (in the control condition). We then gave jurors the admonitions and questions endorsed by the Supreme Court for identifying bias, prior to all of them watching a 32-minute condensed video of a civil trial. Results showed that jurors were simply unable to diagnose their own biases. After excluding jurors admitting or even unsure of their impartiality, jurors exposed to prejudicial news remained significantly more likely to rule against the defendant (odds ratio = 2.4, \( p = .004 \)), and those that did so awarded eight-times larger median damages (\( p = .01 \)), than those exposed to irrelevant news.

Two subsequent experiments, with 751 mock jurors, replicated this finding, but also tested the viability of a remedy suggested by the literature on the “bias blind spot,” namely, that jurors might more accurately diagnose bias in *other* jurors. Jurors asked whether the article they read would bias *others*—rather than themselves—were significantly more likely to say yes. This suggests a simple solution to improve the accuracy of voir dire: stop asking jurors to self-diagnose bias, and instead ask them to diagnose how the pretrial publicity they have seen would bias others. Unfortunately our
test of this possibility suffered from a lack of statistical power, in part because we failed to adequately induce bias.

*Keywords: bias blind spot, self-diagnosis, pretrial publicity, voir dire, bias, introspection*
3. Bias blind spot during voir dire

**Background**

“It is fair to assume that the method we have relied on since the beginning . . . usually identifies bias.”


In late August 2012, a jury in Silicon Valley, California was deliberating on the outcome of a billion dollar patent dispute between two industry giants, Apple and Samsung. The foreman of that jury, Velvin Hogan, was himself an inventor and holder of a patent—one he had litigated, in fact. During voir dire, Mr. Hogan revealed some of these facts in response to the trial judge’s questioning. Mr. Hogan later told the media that he then “expected to be dismissed from the jury.” Nonetheless, the trial judge also asked Mr. Hogan what trial attorneys know as “the magic question”: “[W]ill you be able to decide this case based solely on the evidence that's admitted during the trial?” Mr. Hogan answered affirmatively and was thus seated in the jury, ultimately becoming its foreman. A few weeks later, Mr. Hogan led the jury towards a billion dollar judgment against Samsung, one of the largest in the history of patent law cases.

---

2 Case5:11-cv-01846-LHK Document2013-9, Filed 10/02/12 (Page 3).
3 Case5:11-cv-01846-LHK Document1991-1, Filed 09/21/12 (Page 22).
4 In a series of media interviews, Mr. Hogan said that he wanted the verdict “to send a message to the industry at large that patent infringing is not the right thing to do” and “make sure the message we sent was not just a slap on the wrist.” Case5:11-cv-01846-LHK Document2013, Filed 10/02/12 (Page14) (quoting multiple exhibits).
The fair trial guarantee

The United States Constitution guarantees that litigants within a jury trial receive judgment from an “impartial” panel of jurors. Justice Marshall explained early on that “[t]hose who most prize the institution [i.e., trial by jury], prize it because it furnishes a tribunal which may be expected to be uninfluenced by an undue bias of the mind” (United States v. Burr, 1807, p. 50). Not surprisingly then, failure to obtain an impartial jury is a serious affront to the promise of a “fair” trial. Empaneling even one biased juror can be sufficient structural error to require a new trial (e.g., Ross v. Oklahoma, 1988, p. 85).

The right to an impartial jury is protected by the process of voir dire. In both criminal and civil trials, in both federal and state courts, a certain colloquy is used to screen panels of potential jurors (i.e., venires) and remove unduly biased persons. The judge or attorneys, or both, ask questions to probe whether venirepersons have any feelings or opinions about the litigants, attorneys, facts, or law of the case. If an affirmative answer is given, “it is standard procedure to ask if [the venireperson] can set aside that opinion and decide the case on the basis of the evidence to be presented” and the law as instructed (Suggs & Sales, 1980, p. 246; see also Hannaford-Agor &

[5] The federal criminal defendant’s right to an “impartial” jury is explicitly guaranteed by the Sixth Amendment (“the accused shall enjoy the right to . . . trial by an impartial jury”). Courts have since interpreted juror impartiality as implicit within the notion of due process and, therefore, applicable to all trials. See, for example, Morgan v. Illinois (1992) (“due process alone has long demanded that, if a jury is to be provided the defendant, regardless of whether the Sixth Amendment requires it, the jury must stand impartial and indifferent to the extent commanded by the Sixth Amendment”). Howe (1995) provides a useful review.
3. Bias blind spot during voir dire

Waters, 2004). If this colloquy reveals that a venireperson is and will remain unduly biased, then a challenge for cause may be used to eliminate that person from jury service. If, however, the venireperson self-reports that he or she can be fair and impartial, then—virtually no matter the circumstances behind the possible bias—it is highly unlikely that a challenge for cause will be successful. Both field and experimental studies have found that a venireperson’s self-diagnosis of bias is highly influential, even the strongest factor, in judicial rulings on challenges for cause (e.g., Hannaford-Agor & Waters, 2004; Rose & Diamond, 2008; Kerr, Kramer, Carroll, & Alfini, 1990).

Of course, there are practicalities associated with running a court system. The Supreme Court has explained the caveat that “a litigant is entitled to a fair trial but not a perfect one, for there are no perfect trials” (McDonough Power Equipment, Inc. v. Greenwood, 1984, p. 553). Some imprecision in the screening is tolerated out of necessity. In U.S. v. Casamayor, for example, the Eleventh Circuit upheld a conviction where the jury foreman, due to “inattentiveness,” failed to disclose on a questionnaire his brief relationship with the defendant 23 years prior (1988, p. 1515). And even if more extensive voir dire might uncover subtler forms of bias, there is a tradeoff in terms of court efficiency. Longer voir dire procedures can stall progression on already overburdened court dockets, especially if challenges for cause necessitate bringing in additional venirepersons to seat a complete jury. With these pragmatic constraints in mind, an important normative question is just how much bias we should tolerate in the

6 The attorney might also have a peremptory challenge available to remove a juror.
name of expediency. In other words, on the spectrum of jury impartiality, from absolutely no bias to completely biased, where should we draw the line of tolerance?

This paper does not address where on the spectrum of impartiality is the threshold for when a juror shifts from being merely not perfect to being unfairly biased. We instead examine the primary tool that courts use to decide where a juror sits on that spectrum, namely: simply asking the juror to self-reflect upon whether he or she can be fair and impartial, and then announce to the court his or her self-diagnosis of whether any insurmountable bias exists. Does this colloquy provide any useful information to the litigants and judge, who are together tasked with impaneling an impartial jury? In more general terms, can jurors accurately self-diagnose when bias will alter their judgment?

The psychology behind self-diagnosing bias

A person asked to assess his or her own bias is confronted with a series of difficult psychological obstacles, which suggests the answer might be, “no, jurors cannot accurately self-diagnose.” The self-diagnosis task can be usefully analyzed as part of a more general process of “mental contamination” (Wilson & Brekke, 1994). Wilson and Brekke, in their exposition of the concept, define mental contamination as “the process whereby a person has an unwanted judgment, emotion, or behavior because of mental processing that is unconscious or uncontrollable” (1994, p. 117), and then develop a model outlining what must happen to eradicate mental contamination.

The Wilson and Brekke model highlights at least four separate actions that a person must consciously undertake to successfully eradicate mental contamination—to debias him or herself, and thereby return to a state as if mental contamination never
occurred. A person must: (1) be aware that mental contamination exists; (2) be motivated to correct the bias; (3) be aware of the direction and magnitude of the bias; and (4) be able to adjust his or her response (see also Wilson, Centerbar, & Brekke, 2002). Consider, for example, a potential juror who has been exposed to pretrial publicity that is highly damaging to the defendant in a medical malpractice case. To suppress the biasing influence of this information, the potential juror must first realize the exposure to pretrial publicity has affected his or her judgment. Second, this realization must trigger a desire to counteract the influence of the pretrial publicity. Third, the person must know the direction of the effect (either more or less likely to convict) and its magnitude (e.g., 5% or 85% shift). Finally, he or she must be able to make the necessary correction (e.g., deliberately lower the estimate of guilt by 5% to compensate for a biased 5% increase). It is not trivial to achieve any one of these four steps, much less all four. Let’s examine each more closely.

A person might fail at step one, and not be aware that the pretrial publicity about the defendant has increased the odds of adjudicating against the defendant. A large body of psychological research indicates that awareness of mental contamination can be surprisingly difficult to achieve. One major problem is that people are often ignorant of the cognitive processes by which they form their judgments and decisions (see Nisbett & Wilson, 1977, for a review). Certain outputs of those processes—such as the declarative knowledge activated in short term memory—are accessible to introspection, but not the underlying mechanisms. Vision provides a parallel example: we experience the shapes and colors in our midst, and we can easily report those to others, but it would be an
utterly impossible endeavor to introspect into the cognitive processes interspersed between the moment of photons hitting the retina and visual experience hitting consciousness. This same dynamic is at play in judgment and decision-making. There is, of course, reasoning we can consciously access (see generally Ericsson & Simon, 1993), but there is much we cannot (see Bargh & Morsella, 2008, for a review; see also Bargh & Chartrand, 1999).

A step two failure of motivation would occur if the person were indifferent about her role as a juror or, perhaps more likely, simply disagreed with the judicial instructions that the impact of pretrial publicity constitutes a bias worth eradication. After all, a juror is not a member of the legal culture, which is indoctrinated through the process of law school to believe that the epistemic sterility of a trial under the Rules of Evidence is the optimal way to decide an important question. Instead, jurors may feel that they should rely on all available information, and even their feelings about the parties, just as the jurors do in the rest of their lives. Research on jury nullification suggests that jurors are, in certain limited circumstances, willing to disregard the law in order to secure what they perceive is the merited outcome (Devine, 2012, pp. 68–71).

Even if aware and motivated, it is unlikely that a juror would have any insight into how much bias the pretrial exposure caused (a failure of step three). This is evidenced by the fact that when participants are asked to correct for an explicit bias, they usually undercompensate or overcompensate, and only very rarely hit the nail on the head. Consider the anchoring effect, wherein a numerical judgment is biased by previous consideration of an irrelevant number. In the classic study by Kahneman and Tversky, for
example, subjects were asked to estimate the number of African countries in the United Nations (1974). Answers were highly correlated with whatever number was first randomly generated from a wheel of fortune. Wilson et al. used this design, but also: explicitly warned subjects that anchors would bias their judgments; provided an example; and admonished them to “please be careful not to have this contamination effect happen to you” (Wilson, Houston, Etling, & Brekke, 1996). Clearly aware of the possibility of bias, subjects attempted to compensate, but the magnitude of their shift was totally off target. In this case they severely underestimated how much they needed to adjust, presumably because they likewise underestimated just how strong an effect anchors can exert.

A failure at the final (fourth) step entails an inability to correct for a bias that is precisely known and unwanted. Some bells cannot be un-rung. Explaining what this could mean is easiest by way of example. Rozin, Millman, and Nemeroff (1986) asked people to engage in tasks such as drinking apple juice from a brand new bed pan or eating chocolate shaped like dog feces. A strong disgust response was elicited that subjects simply could not shake, despite knowing the bed pan was sterile or the chocolate sweet. Rozin later recalled subjects “often laugh[ed] at themselves or almost apologize[d],” realizing it was a reaction they wish they could overcome—but they just couldn’t (Rozin, Grant, Weinberg, & Parker, 2007). A juror who heard horrible facts about a defendant might similarly be unable to shake a feeling of moral disgust.

Given what we know about the limits of human cognition, and the problem of mental contamination in particular, it would not be surprising if jurors occasionally or
even usually fail the task of self-diagnosing their own biases. Indeed, one of the authors recently presented an abstract of this study at a conference, with the title, “Can jurors self-diagnosis bias?” A psychologist briskly walking past the poster—apparently in some sort of hurry—still found the time to make eye contact, raise his finger, and proclaim, “No!” before vanishing around the corner. (One needs a good hindsight bias joke for such moments.) But what do legal actors—in particular the judges policing voir dire—think about this ultimately empirical question? How much faith do they place in jurors’ ability to self-diagnose bias? The answer seems to be: A lot.

The Skilling test for bias: Juror self-diagnosis

The American judicial system weighs a juror’s self-diagnosis during voir dire very heavily, displaying an assumption that, as a factual matter, jurors can do this self-assessment accurately. This is evident in both written opinions—explicitly saying as such—and as an empirical matter, if one observes how the voir dire process unfolds in courtrooms across the country.

A search through the case law uncovers countless quotes wherein judges display their presumption that juror self-diagnosis provides valuably accurate information. Appeals arguing juror bias are swatted down left and right with the observation that jurors denied bias and, therefore, bias must not have existed (see e.g., Magna Trust Co. v. Illinois Cent. R. Co.(2000), emphasizing that the juror “repeatedly stated that she could be fair and impartial”; or Sawyer v. Southwest Airlines Co. (2005), observing that “[d]uring voir dire, these prospective jurors stated that they could follow the court’s instructions and render a fair verdict”). Courts hold out self-diagnosis as the way—
maybe the only way—for definitively rooting out bias (see e.g., *Com. of Pennsylvania v. Sandusky* (2012), asserting that “the answer to whether a juror can be fair and impartial, despite the myriad of influences to which he or she may be exposed, cannot be known until the juror is actually asked”). And they are unmoved by insinuations that self-diagnosis is given too much weight (see e.g., *State v. Addison* (2010, p. 58), “reject[ing] the defendant’s argument that our reasoning . . . is flawed because it ‘relies too heavily on jurors’ assurances that they can be fair’”). Indeed some courts (at least older ones), to the extent they are suspicious of self-diagnosis at all, doubt it in the opposite direction; they worry jurors might *feign* bias in order to dodge jury duty (e.g., *Reynolds v. United States*, 1879, noting “[their] experience [] that jurors not unfrequently [sic] seek to excuse themselves on the ground of having formed an opinion, when, on examination, it turns out that no real disqualification exists”). Are these rogue judges; quotes cherry-picked to illustrate a point?

One must look no further than the 2010 case *Skilling v. U.S.* for a recent incantation of the magic question colloquy, and its endorsement by the highest court of the land. The U.S. Supreme Court considered former Enron executive Jeffrey Skilling and his (denied) motion for a change of venue out of Houston—the city where his crumpling company had been headquartered, and thus where several thousand people had lost their jobs and fortunes. Skilling argued that the trial “never should have proceeded in Houston” due to “the community passion aroused by Enron's collapse and the vitriolic media treatment” (*Skilling v. US*, 2010, p. 2912). Even if had been possible to pick the
handful of impartial jurors from the otherwise vitriolic venire, he continued, “[t]he truncated voir dire . . . did almost nothing to weed out prejudices” (2010, p. 2912).

The voir dire in Skilling’s trial was extensive, although led by the judge (who refused to allow the attorneys to question the jurors). The trial judge had solemnly instructed the potential jurors that, “The bottom line is that we want . . . jurors who . . . will faithfully, conscientiously[,] and impartially serve if selected” (2010, p. 2910).

“[E]ach of you,” the court further explained, “needs to be absolutely sure that your decisions concerning the facts will be based only on the evidence that you hear and read in this courtroom” (2010, p. 2911). In all, two potential jurors were excused after stating they could not be impartial, one out of five was removed for cause at the government’s request, and three out of nine were removed for cause at the defendant’s request.

Skilling was convicted of 19 counts and acquitted of 9 counts at trial. The Fifth Circuit reversed the convictions, holding that the “magnitude and negative tone of media attention directed at Enron” created a presumption of bias (2010, p. 2916). The Supreme Court, however, reinstated the convictions, holding that Skilling was not denied a fair trial and that he did not prove that the jury was biased.7

On the question of whether the jury was in fact biased by pretrial publicity, the Court expressed great deference for the trial court’s determination, which was itself

7 The majority opinion cited four criteria when a change of motion should be granted to ensure a fair trial: (1) the size of the community in which the trial takes place; (2) the content of the trial coverage, whether or not any confessions or other “blatantly prejudicial information” is in the news; (3) the amount of time between the trial and the initial news coverage of the crime; and (4) whether or not the jury convicted the defendant of all counts against him (Skilling v. US, 2010, pp. 2915–2916).
3. Bias blind spot during voir dire

Based on the jurors’ own self-assessments and assurances of impartiality. As the Supreme Court explained matter of factly, “in response to the question whether ‘any opinion [they] may have formed regarding Enron or [Skilling] [would] prevent’ their impartial consideration of the evidence at trial, every juror—despite options to mark ‘yes’ or ‘unsure’—instead checked ‘no’” (2010, p. 2921). When rebutting Justice Sotomayor’s dissent, the Skilling majority again cited back to these juror self-diagnoses (2010, pp. 2919–2921, e.g., emphasizing again that jurors answered “yes” when asked if they “ha[d] an opinion about ... Skilling”). And yet a third time, the majority relied on the idea that, “all of Skilling's jurors had already affirmed on their questionnaires that they would have no trouble basing a verdict only on the evidence at trial,” and noted that the trial court had nonetheless removed one such juror, who said he could “abide by the law” (2010, pp. 2915–2916, n. 30).

To give a flavor of the deference, consider that one juror “stated that ‘greed on Enron’s part’ triggered the company's bankruptcy and that corporate executives, driven by avarice, ‘walk a line that stretches sometimes the legality of something’” (2010, p. 2924). The Supreme Court nonetheless found it appropriate to seat the juror because “he also asserted that he could be fair and require the government to prove its case” and because the trial judge had “looked [Juror 11] in the eye and . . . heard all his [answers], [and the trial judge] found his assertions of impartiality credible” (2010, p. 2924).

The Skilling case makes clear that, going forward, courts should and will give considerable weight to the self-professed neutrality of jurors, although with the
potential to disregard such professions if a stern “look in the eye” suggests that they are not credible. State courts likewise follow the procedure endorsed by *Skilling*, and some have statutes requiring that “a judge must inquire whether a prospective juror has expressed or formed an opinion on a case or is aware of any bias or prejudice” (*Commonwealth v. Entwistle*, 2012, p. *13).

Of course there are caveats, and one can find notable dissenting voices. Several opinions emphasize that the self-diagnosis inquiry is no “magic question,” the answer to which trumps all other context the judge might consider (e.g., *Montgomery v. Com.*, 1991, explaining that “[t]here is no ‘magic’ in the ‘magic question.’ It is just another question where the answer may have some bearing on deciding whether a particular juror is disqualified by bias or prejudice, from whatever source”). And jurists have at selected times hinted at possible psychological foibles (e.g., *Irvin v. Dowd*, 1961, with Justice Clark musing that “[n]o doubt each juror was sincere when he said that he would be fair and impartial to petitioner, but the psychological impact requiring such a declaration before one’s fellows is often its father. Where so many, so many times, admitted prejudice, such a statement of impartiality can be given little weight”).

Notably though, even these caveating and dissenting judges seem to place great weight on self-diagnosis in general. They simply stress that other contextual factors, if truly overwhelming, outweigh the self-diagnosis in particular instances. The judge on the bench is expected to sift through the self-diagnoses proffered from the venire and screen out the wheat from the chaff. Do judges in fact sort through the juror proclamations, only giving weight to the worthy select?
The empirical evidence reveals that judges place considerable, often overwhelming, weight on self-diagnosis. A recent study of California cases found that once a potential source of bias was identified, if a juror nonetheless said that he or she could be fair, it made him or her 71% less likely to be dismissed for cause, all other things being equal (Hannaford-Agor & Waters, 2004). And once the trial judge has made the determination that a juror can be fair, it is virtually unreviewable.

**General Method**

**Analytic strategy**

Courts assume that “the method we have relied on since the beginning” (i.e., voir dire, with its reliance on self-diagnosis) “usually identifies bias” (Patton v. Yount, 1984, p. 1038). To test the assumption that self-diagnosis is accurate—an empirical claim—we must specify counterfactuals for the hypothesis that voir dire is not effective at rooting out bias and the hypotheses that it is effective. We ask whether a juror, whose impartiality has been questioned due to exposure to a potentially biasing factor, can nonetheless decide the case the same as if he or she had never been exposed to the biasing factor. We do not require the juror to be ignorant of potentially biasing factors; it is sufficient if he or she can set aside those factors and decide the case in an unbiased way, which is all the law expects (see e.g., *Skilling v. US*, 2010, noting that “juror impartiality does not require ignorance,” and jurors “need not enter the box with empty heads in order to determine the facts impartially”).

We implement this analytic strategy empirically by way of a between-subjects design with two steps, one experimental and one statistical. Before describing the steps,
Note that all jurors, regardless of condition, hear the exact same case. They also answer, per the colloquy endorsed in *Skilling*, whether they can be fair and impartial. And all jurors proceed to render a verdict regardless of their answer. Now for the two steps.

In the first step, we deliberately bias half of subjects. In this instance we have subjects in the “treatment” group read an inadmissible newspaper article that is highly damaging to the defendant, while subjects in the “control” group read an irrelevant article. We know from prior research that such pretrial publicity can be highly biasing, in that treatment subjects become substantially more likely to find against the defendant than control subjects (Steblay, Besirevic, Fulero, & Jimenez-Lorente, 1999, provide a meta-analytic review). We aim to replicate that finding here, but only to create a pool of biased jurors to study in terms of self-diagnostic accuracy. Note the state play at this moment. If the pretrial publicity stimuli is effective, then the control juror pool will find against the defendant at some rate $G$, while the treatment juror pool will find against the defendant at a higher rate, $G + b$, the difference $+ b$ due those jurors who were sufficiently biased by the pretrial publicity. We measure this by comparing the control and treatment groups, inclusive of all subjects; we expect $G_{\text{treat}} - G_{\text{control}} = b$.

In the second step, we remove subjects who concede bias from the dataset. We then re-examine the verdict rates across conditions, using only those subjects who insist they are fair and impartial—those who “pass” the *Skilling* screen, so to speak, and thus are fit for jury duty. Label these screened pools $G_{\text{treat, screened}}$ and $G_{\text{control, screened}}$. The key test is then $G_{\text{treat, screened}}$ minus $G_{\text{control, screened}}$. If people are able to self-diagnose accurately, then the effect of statistically removing those jurors who concede bias would
be to remove the biased verdicts captured by the $+ b$ term, which would mean the differential verdict rate seen before the screen (i.e., $G_{treat} - G_{control} = b$) is no longer seen after the screen (i.e., $G_{treat\_screened} - G_{control\_screened} = 0$). For example, if 30% of jurors impose liability in the control condition, while 50% impose liability in the treatment condition (i.e., $G_{treat} - G_{control} = 50\% - 30\% = 20\%$), then the Skilling protocol will succeed in its goal of providing a fair trial if the screened treatment condition drops down to a 30% conviction rate (i.e., $G_{treat\_screened} - G_{control\_screened} = 30\% - 30\% = 0\%$). Note this same logic applies for any other outcome variables, such as monetary damages.  

The biasing stimuli: Pretrial publicity (PTP)

There are a wide variety of biases and prejudices that might inflict jurors, any of which could be studied in terms of self-diagnostic accuracy. (Indeed an important question is whether jurors might be better or worse at diagnosing different biases). Here we selected bias due to pretrial publicity (PTP) for primarily methodological reasons: PTP is relatively easy to manipulate experimentally, and there is an expansive literature demonstrating it can have powerful effects (Steblay et al., 1999). But PTP is

---

8 Note the group-level bias could reflect one of two (or a mixture of both) individual-level effects. One possibility is that a subset of jurors in the treatment condition unavoidably succumb to bias such that their verdict changes, while others altogether avoid such mental contamination. The Skilling screen succeeds to the extent the former group accurately identifies itself as biased, while the latter group accurately identifies itself as fair and impartial. The second possibility is that everyone in the treatment condition is biased to some degree (e.g., their guilt assessment shifts 5%), with the differential conviction rate between control and treatment groups reflecting only that subset of biased persons for whom this shift transverses the threshold for a verdict change (e.g., a verdict shift in the likelihood of guilt from 10% to 15% would remain a not guilty verdict; a shift from 90% to 95%, however, might cause a change from not guilty to guilty). Under this conception, the Skilling screen succeeds to the extent the subset surpassing the verdict change threshold accurately identifies itself as biased; those at the extremes could be inaccurate, since their verdict is unchanged regardless. Note that if jurors accurately self-identify as a function of a percentage-shift (e.g., they realize their verdict likelihood will shift 5%), but do not know or use the base rate, then some will remove themselves who binary verdict would not have changed.
also an important problem in its own right. PTP is, naturally, most problematic in highly publicized trials, which means that the taint of PTP can disproportionately impact perceptions of judicial legitimacy. The influence of PTP is also a growing problem, due to social media and the 24-hour news cycle (see e.g., Morrison, 2011, discussing jurors’ use of legal and factual information on the internet).

As it happens, the history of voir dire itself traces back to PTP, in particular the famous trial of Aaron Burr (see Dyke, 1975, recounting how “[t]he media of the day described the feud between Jefferson and Burr in detail, the citizenry chose sides, and the difficulties in selecting an impartial jury increased”). Interestingly, Chief Justice John Marshall recognized early on that the court was grappling with a difficult psychological issue, one wherein a juror can deceive him or herself:

The relationship may be remote; the person may never have seen the party; he may declare that he feels no prejudice in the case; and yet the law cautiously incapacitates him from serving on the jury because it suspects prejudice, because in general persons in a similar situation would feel prejudice. He will listen with more favor to that testimony which confirms, than to that which would change his opinion; it is not to be expected that he will weigh evidence or argument as fairly as a man whose judgment is not made up in the case.

Notably, Justice Marshall recognized that the pretrial publicity created a “suspicion” of prejudice or “bias,” an empirical claim that must be resolved by the judge. Marshall said that the trial court should question such jurors to decide whether they are “capable of hearing fairly, and of deciding impartially, on the testimony which may be offered to them, or as possessing minds in a situation to struggle against the conviction which that testimony might be calculated to produce.” Marshall himself expressed some doubt about the value of such a colloquy, since there may be prejudice even
where the juror “declares that he feels” none. Marshall’s opinion was extremely
influential for both the state and federal courts, who adopted the practice of questioning
jurors, a practice that did not and still does not exist across the Atlantic.

Part I: The inability to self-diagnose bias due to PTP

In Part I, we ask whether courts are justified in placing such reliance on self-
diagnoses. The first two experiments, one with a sample of law students and a second
with a larger and more diverse national sample online, tested the ability of jurors to self-
diagnose bias. Both experiments used the same design, stimuli, and instrument. In
particular, we used a 2×1 between-subjects experimental design, wherein subjects were
exposed to either irrelevant or prejudicial pretrial publicity concerning the defendant.
The analytic approach outlined above was then used to assess the accuracy by which
jurors could accurately self-diagnose bias.

Experiment 1: Law student self-diagnosis

**Method**

**Participants**

A convenience sample of first year law students (N = 74) were randomly assigned
to unequally sized control (n = 30) and treatment (n = 44) groups. Demographic
variables were not collected. We assigned more to the treatment group assuming more
would disqualify themselves during voir dire, and thus a larger sample would be
necessary to ensure the sample sizes were roughly equivalent for use in post-screen
analyses.
3. Bias blind spot during voir dire

Procedure

The analytic approach outlined above was used. Across both conditions, all subjects engaged in the same set of voir dire questioning and then, regardless of their self-diagnoses, proceeded to render a verdict on a medical malpractice trial, delivered via a 32-minute video. Jurors conceding bias were then statistically removed in order to operationalize the voir dire screen prescribed by *Skilling*, and then verdict rates were compared again.

Biasing stimuli

The pretrial publicity stimuli were based on either of two abridged articles from the Kansas City Star, each about 1300 words in length. Study materials are available upon request; the original articles were by Alan Bayley (2011) and Diane Stafford (2011). The control article discussed a topic irrelevant to the trial, namely, employer incentive programs for preventative health maintenance. The treatment article was modified to actually name the defendant in the trial, Dr. John Dennis, and was highly damaging: it discussed a prior case of medical malpractice, a history of prior malpractice claims and settlements much higher than the national average, and the effects of the malpractice on the other injured patients—all thereby painting Dr. Dennis in a negative light. The article suggested that most medical malpractice is due to a few bad doctors who need to be taken out of the system.

Courts use a rigorous set of evidentiary rules to carefully delimit the information available to jurors, so as to ensure that they base their decision only on proper evidence. The article used as PTP in the treatment condition included numerous facts inadmissible
under the federal rules of evidence (FRE). For example, the information regarding Dr. Dennis’ prior incidents of malpractice would be inadmissible pursuant to FRE 404 (“Evidence of a crime, wrong, or other act is not admissible to prove a person’s character in order to show that on a particular occasion the person acted in accordance with the character”). The statements made by those who had previously sued Dr. Dennis would most likely be irrelevant under FRE 401 and are hearsay under FRE 801. The statistics quoted in the article are also hearsay. Thus, any effect of the treatment article would represent clear error, as that information should be strictly excluded from consideration in trial.

**Operationalizing the Supreme Court’s voir dire protocol**

Subjects in both conditions were given written admonitions from a mock judge about the need to be impartial, and they were asked a series of questions about whether they could be impartial. These questions were directly drawn from the judge-juror colloquy affirmed in *Skilling v. U.S.* (2010). Specifically, the opening and mock voir dire went as follows:

“You are called into jury duty. After waiting in the jury commissioner’s office, you are ushered into the courtroom. The judge calls you to the bench individually, and he says:

You have been called to potentially be a juror in a medical malpractice case involving Mr. Andrew Stevens, as the plaintiff, suing Dr. John Dennis as the defendant. It is important for Mr. Stevens, Dr. Dennis, and for our legal system that the jurors be fair and impartial. Jurors must decide the case based only on the evidence presented during the trial, and not based on any prejudices, biases, preconceived ideas, or extraneous information.

The bottom line is that we want jurors who will faithfully, conscientiously, and impartially serve if selected. Each of you needs to be absolutely sure
that your decisions concerning the facts will be based only on the evidence that you hear and read in this courtroom.

Unfortunately, I understand that some of you may have seen some news items about one or more of the parties in this case, or may have negative opinions about doctors, patients who sue, or the healthcare system at large. This fact alone does not automatically disqualify you from hearing this case however. You have a duty to perform your civic duty as a juror, if you can be fair and impartial in doing so.

Therefore, I am going to ask you a few questions. And, there are no right or wrong answers to the questions.”

Subjects then answered the following questions, with potential answers of “yes,” “no,” or “unsure”:

1. “Did you read a news article about Dr. John Dennis?”
2. “Do you have an opinion about Dr. John Dennis?”
3. “Would any opinion you have prevent your impartial consideration of the evidence at trial?”; and
4. “Could you base a verdict only on the evidence at trial?”

The first question was just a manipulation check. To simulate the screening process endorsed by the Supreme Court in *Skilling*, we used questions #3 and #4 for analysis. A subject answering “yes” or “unsure” to #3, saying that her opinions would or may prevent impartial consideration of the evidence, would be disqualified. And a juror answering “no” or “unsure” to #4, saying she could not base a verdict only on the evidence at trial, would be excluded from the jury. We understand that this is the test
endorsed by the Supreme Court in *Skilling*, in accordance with prior cases going back to Justice Marshall’s opinion in *Burr*.\(^9\)

**Trial stimulus**

All subjects watched a 32-minute medical malpractice trial video that included opening statements from the plaintiff’s and defendant’s attorneys, testimony from expert witnesses about the standard of care in the case, cross-examination of both experts, closing statements from the plaintiff’s and defendant’s attorneys, and finally jury instructions from the judge. This video was developed by real physicians serving as writers of the medical scenario and serving as actors for the expert witnesses, along with an experienced arbitrator consulting on the jury instructions and serving as the judge. Two of the co-authors served as attorneys. Thus, although condensed, the video had a high degree of verisimilitude.

The scenario in the video concerned the failure of a primary care physician to diagnose a possible case of lumbar radiculopathy and refer the patient to imaging, which allegedly would have allowed timely surgery and avoidance of the permanent disability that the patient now suffers. The primary dispute concerned whether the physician-defendant met the standard of care when, instead of ordering imaging, he simply instructed the patient to take over-the-counter medications and return if the pain got worse. An actor posing as a judge provided jury instructions, which were based on the standard templates used in Arizona medical malpractice cases.

\(^9\) For these purposes, we also treated “unsure” as a disqualifying answer. Robustness checks revealed that allowing “unsure” respondents to remain does not change the results.
3. Bias blind spot during voir dire

**Outcome measures**

Subjects rendered individual judgments, responding “yes” or “no” to the prompt: “Based on the instructions provided by the judge in the video, do you believe that the Plaintiff has proved, by the greater weight of the evidence, that the Defendant committed medical negligence?” They also rated the case on a 6-point Likert scale from (1) “certainly not medical negligence” to (6) “certainly medical negligence.” Jurors who found negligence also awarded non-economic damages for “pain and suffering,” which had been defined by the judge’s instructions. (Economic damages were not discussed in the trial, and were presumed to be stipulated by the parties).

**Results**

**The effect of pretrial publicity**

The treatment PTP biased verdicts as expected. Referring to the first major row of Table 1 (all jurors), only 13% (4 of 30) of the jurors in the control group voted against the defendant. Of the jurors who exposed to negative PTP in the treatment condition, on the other hand, 32% (14 of 44) voted against the defendant. See also Figure 2. This 19% increase, 95% CI [0%, 37%], in verdicts against the defendant indicates that, as hypothesized, exposure to pretrial publicity biases jurors. The effect is substantial; with an odds ratio of 3.03, 95% CI [0.9, 10], exposure to prejudicial publicity more than doubled the odds of a verdict against the defendant. Statistically speaking the result is marginally significant, \( \chi^2 (1) = 3.31, p = .07 \), most likely reflecting the small sample size.\(^{10}\)

\(^{10}\) Textbook statistics generally recommend that all cell sizes within a chi-square test contain at least 5 observations. The reason is that estimates of the chi-square distribution rely on large-sample theory, an
3. Bias blind spot during voir dire

Figure 2. Verdict rates by negative PTP manipulation, pre and post Skilling screen, in the Experiment 1 (left) and Experiment 2 (right). In both cases jurors completely failed to accurately self-diagnose bias.

The effect on awarded damages would be the variable of most interest to players in civil litigation. On this point, we find a clear, significant effect: jurors in the treatment (“exposed”) condition imposed nearly nine times as much damages for pain and suffering as the jurors in the control condition (mean and 5% trimmed mean scores, respectively, of $98,500 and $81,286 versus $10,600 and $10,600; $U = 832$, $z = 2.42$, $p = .015$). Defense verdicts were counted as having damage awards of zero dollars, and since a preponderance of verdicts were for the defense, the median award in both

assumption that is possibly violated when cells are smaller than 5. There is, however, debate about whether and when small cell sizes undermine the chi-square test (see e.g., Ruxton & Neuhäuser, 2010). For our purposes with this initial pilot study, the $p = .069$ statement is suggests enough reliability to motivate the second experiment, without need to delve into this chi-square debate.
conditions was zero. However, even when defense verdicts are excluded, we still see a robust shift in median damages awards, $33,000 in the unexposed control condition versus $300,000 in the exposed treatment condition \( p = .05 \). Our intervention succeeded in creating a bias.

**The accuracy of jurors' self-diagnoses**

The more interesting question is whether the courts can use juror self-assessments to eliminate that bias, and thus secure the defendant’s right to a fair trial. Of the 44 respondents in the treatment group, five (11%) thought that they could not be impartial, 15 (34%) indicated that they were unsure, and 24 (55%) indicated that they could be impartial. Suppose that all those who thought the pretrial information would prevent them from being impartial, or were unsure about that point, were removed from the jury as prescribed by the Supreme Court in *Skilling*.\(^{11}\) If the *Skilling* procedure is effective, then the remaining jurors (which we label “self-screened”) should vote similarly as do those in the control condition. This is the analytic strategy we take (see p. 71 above for more detail).

The evidence of bias persists despite the *Skilling* screen. Of the allegedly impartial subgroup of the treatment condition \( n = 24 \), seven (29%) voted against the defendant. Compared with the control group \( n = 24 \), where only three (13%) voted

---

\(^{11}\) For our experimental purposes, we will exclude such jurors from both the control and treatment conditions, since jurors may have other bases for doubting their own partiality (e.g., a personal experience with malpractice), which we did not explore in our truncated voir dire. In a real trial, jurors may not be asked such a question with particularity, unless there were prima facie concerns about bias (which do not arise in our control condition). See e.g., *Mu’Min*, 500 U.S. at 420 (discussing trial judge’s questioning procedure).
against the defendant, it seems that the *Skilling* protocol failed—subjects exposed to prejudicial information, despite claiming not to be affected by it, were 2.88, 95% CI [0.7, 12], times as likely to adjudicate that malpractice occurred than were subjects exposed only to irrelevant information. As before a problem of small sample size appears likely—indeed it is exacerbated by the Skilling exclusions of subjects—with this difference failing to achieve traditional statistical difference, $\chi^2 (1) = 2.02, p = .16$.

Examination of pain and suffering awards, where mean and 5% trimmed mean scores awards of $87,542 versus $15,000 were observed for the treatment and control conditions, respectively, now shows a failure to reach traditional significance, $U = 341.5, z = 1.55, p = .12$. But this too likely reflects a problem with sample size, given that the mean award in the treatment condition was over six times as large as that within the control condition. Even excluding zeros, we see a quintupling of median damages awarded, from $60,000 to $300,000 ($p = .27$), though again short of statistical significance on this small sample.

Results on the 6-point scale as to the “certainty” of negligence (not shown) mimicked the above results. Treatment subjects exposed to prejudicial publicity were significantly more certain than those in the control condition exposed to irrelevant publicity (3.3 vs. 2.8, respectively), $t(72) = -2.47, p = .02$. And this remained true (albeit marginally, by traditional significance standards) even when excluding those who admitted or were unsure of bias as *Skilling* prescribes (for that subsample, scores were 3.2 vs. 2.8, respectively, $t(46) = -1.93, p = .06$). These suggestive results motivated a second study with greater statistical power, and a more representative sample.
3. Bias blind spot during voir dire

Table 1. Convenience sample. In a randomized experiment with 74 law students, those exposed to a news article that negatively portrayed the defendant voted against him more often (“liability votes”), and awarded higher damages (even after excluding zeros). Prior to viewing the trial, jurors were asked, “Would any opinion you have prevent your impartial consideration of the evidence at trial?” and “Could you base a verdict only on the evidence at trial?” Exclusion of jurors who said ‘no’ or ‘unsure’ (the “self-screen”) failed to remove this bias. Statistical power was limited however.

<table>
<thead>
<tr>
<th>Jurors</th>
<th>N</th>
<th>Condition</th>
<th>Can be Impartial</th>
<th>Screened Out</th>
<th>Liability Votes</th>
<th>Bias Gap</th>
<th>Odds Ratio (.95CI)</th>
<th>Mean Damages</th>
<th>5% Trim</th>
<th>SD</th>
<th>Nonzero Median</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>74</td>
<td>Unexposed</td>
<td>30</td>
<td>24 (80%)</td>
<td>6 (20%)</td>
<td>0 (0%)</td>
<td>4 (13%)</td>
<td>19% .07</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exposed</td>
<td>44</td>
<td>24 (55%)</td>
<td>15 (34%)</td>
<td>5 (11%)</td>
<td>14 (32%)</td>
<td>19% .07</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-screened</td>
<td>48</td>
<td>Unexposed</td>
<td>24</td>
<td>24 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>3 (13%)</td>
<td>16% .16</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exposed</td>
<td>24</td>
<td>24 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>7 (29%)</td>
<td>16% .16</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. National sample. In a randomized experiment 174 jury-eligible persons recruited online, those exposed to a news article that negatively portrayed the defendant voted against him more often (“liability votes”), and awarded higher damages (even after excluding zeros). Prior to viewing the trial, jurors were asked, “Would any opinion you have prevent your impartial consideration of the evidence at trial?” and “Could you base a verdict only on the evidence at trial?” Exclusion of jurors who said ‘no’ or ‘unsure’ (the “self-screen”) failed to remove this bias.

<table>
<thead>
<tr>
<th>Jurors</th>
<th>N</th>
<th>Condition</th>
<th>Can be Impartial</th>
<th>Screened Out</th>
<th>Liability Votes</th>
<th>Bias Gap</th>
<th>Odds Ratio (.95CI)</th>
<th>Mean Damages</th>
<th>5% Trim</th>
<th>SD</th>
<th>Nonzero Median</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>174</td>
<td>unexposed</td>
<td>65</td>
<td>59 (91%)</td>
<td>6 (9%)</td>
<td>0 (0%)</td>
<td>23 (35%)</td>
<td>17% .03</td>
<td>.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>exposed</td>
<td>9</td>
<td>95 (87%)</td>
<td>9 (8%)</td>
<td>5 (5%)</td>
<td>57 (52%)</td>
<td>17% .03</td>
<td>.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-screened</td>
<td>15</td>
<td>unexposed</td>
<td>59</td>
<td>59 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>20 (34%)</td>
<td>19% .02</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>exposed</td>
<td>95</td>
<td>95 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>50 (53%)</td>
<td>19% .02</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment 2: National sample self-diagnosis

We aim here to replicate the finding from Experiment One, also using a larger and more diverse sample.

Method

Participants

Jury-eligible adults (viz., United States residents, 18 years or older, who could read, write, and speak English) were recruited from Amazon Mechanical Turk (“MTurk”) in May and July of 2012. Subjects were compensated four dollars, with an opportunity of bonus pay for respondents who scored highest on tests of recall. (This was to incentivize attention and effort within the experimental task). All subjects consented in accordance with IRB requirements.

Two hundred and sixty-four persons proceeded past the initial informed consent webpage into the experiment. Of these, 64 persons exited the study before completion, constituting an attrition rate of 24%. An additional 26 persons were excluded for failure to comply with the task. In particular, they finished the experiment, which entails a 32-minute video and several pages of questions, in the impossibly fast time of 34 minutes or less. The final sample thus includes 174 subjects. Demographic variables of sex,

12 The length of the video – 32 minutes – is an obvious and objective threshold for exclusion. Setting a threshold of greater length requires an estimate of how quickly it would be possible to read and answer the task materials, which in turn requires a judgment that risks the possibility of a false positive (wrongly excluding a subject who happens to work quickly). Examination of the data revealed a sharp gap, with two distributions. Those labeled a “cheater” almost all finished under the 32-minute mark; only two persons took longer, and each were below 34 minutes. The tail of the other distribution, which represents the quickest “non-cheating” score, was 44 minutes.
age, education, and gender were examined to explore whether characteristics of the person predicted whether he or she dropped out or cheated. None had predictive power, and the excluded group demographically resembled the final sample.

As can be seen in Table 3, the overall sample roughly resembled U.S. Census data. Participants were on average more educated and younger, although neither characteristic predicted verdict (see regression analyses in Attachment B). Especially relative to jury research using convenience college samples, our sample constitutes a respectable level of representativeness, and allows reasonable inferences about the jury pool at large. Demographic variations were fairly well distributed across the experimental conditions, showing that randomization succeeded. Subsequent analyses thus proceed using condition-splits, proportions, and central tendencies for ease of communication. Regression tables are presented in the Attachment B.\textsuperscript{13}

\textsuperscript{13} Attachment B reports the results of regression analyses of models testing condition as a predictor, as well as nested models testing for any additional variance accounted for by self-diagnosis, psychological, and demographic constructs. The regression analyses reveal conclusions that are equivalent to the chi-square analyses discussed in the body. In particular, a model regressing verdict on condition is significant, $\chi^2 (1) = 4.74, p = .029$, and remains significant after exclusion of jurors who fail the \textit{Skilling} screening, $\chi^2 (1) = 5.22, p = .022$. Nested model comparisons revealed that self-diagnosis, NFC, and CRT added no significant predictive power, $\chi^2 (3) = 1.06, p = .787$, nor did any of the demographic variables (race, sex, education, age), $\chi^2 (4) = 0.34, p = .987$. 

Table 3. Demographics by condition. The overall sample from Amazon Mechanical Turk roughly resembled U.S. census data, although it was on average younger and more educated. The control and treatment conditions were relatively similar in their demographic distributions.

<table>
<thead>
<tr>
<th>Education</th>
<th>Control (n = 65)</th>
<th>Treatment (n = 109)</th>
<th>Total (N = 174)</th>
<th>U.S. Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; HS Diploma/GED</td>
<td>3%</td>
<td>1%</td>
<td>2%</td>
<td>18%</td>
</tr>
<tr>
<td>HS Diploma/GED</td>
<td>15%</td>
<td>9%</td>
<td>12%</td>
<td>30%</td>
</tr>
<tr>
<td>Some College/Assoc.</td>
<td>49%</td>
<td>48%</td>
<td>48%</td>
<td>27%</td>
</tr>
<tr>
<td>College Grad</td>
<td>25%</td>
<td>30%</td>
<td>28%</td>
<td>17%</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>8%</td>
<td>12%</td>
<td>10%</td>
<td>10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>42%</td>
<td>45%</td>
<td>44%</td>
<td>49%</td>
</tr>
<tr>
<td>Female</td>
<td>58%</td>
<td>55%</td>
<td>56%</td>
<td>51%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Groups</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18-24</td>
<td>17%</td>
<td>28%</td>
<td>23%</td>
<td>13%</td>
</tr>
<tr>
<td>25-34</td>
<td>54%</td>
<td>31%</td>
<td>40%</td>
<td>18%</td>
</tr>
<tr>
<td>35-44</td>
<td>17%</td>
<td>20%</td>
<td>19%</td>
<td>19%</td>
</tr>
<tr>
<td>45-59</td>
<td>11%</td>
<td>18%</td>
<td>15%</td>
<td>27%</td>
</tr>
<tr>
<td>60+</td>
<td>1%</td>
<td>3%</td>
<td>3%</td>
<td>23%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>80%</td>
<td>77%</td>
<td>78%</td>
<td>74%</td>
</tr>
<tr>
<td>Non-White</td>
<td>20%</td>
<td>23%</td>
<td>22%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Procedure

The same procedure as in Experiment One was used, although demographic information and cognitive scales were also collected.
Results

The effect of pretrial publicity

As with the pilot experiment, our analyses proceeded in two steps. We first asked whether negative PTP had a biasing effect by examining if, as hypothesized, subjects exposed to it were more likely to find that medical malpractice occurred, to award larger monetary damages, and to be more certain of their verdict than those subjects exposed to irrelevant publicity. Second, we examined whether, as hypothesized, the screen proposed by the Supreme Court, most recently in Skilling, was sufficient to remove the biasing effect of prejudicial publicity observed in the first step.

Exposure to prejudicial pretrial publicity did, as hypothesized, significantly bias jurors. As seen in Table 2 (“all jurors” row), only 35% of jurors in the control condition found medical negligence, but this percentage increased to 52% amongst those jurors exposed to prejudicial publicity—an absolute increase of 17%, 95% CI [2%, 32%], $\chi^2 (1) = 4.687$, $p = .030$, $\phi = .16$. This amounts to an odds ratio of 2.00, 95% CI [1.1, 3.8], meaning that exposure to prejudicial pretrial publicity doubled the odds of a verdict against the defendant. It is worth emphasizing again that the trial stimulus itself was exactly identical across conditions, and thus a strong inference can be made that the negative PTP was the casual source of this change in verdict rates.

The effect on pain and suffering damages, the ultimate variable of interest to practicing litigators, was more dramatic. The median and 5% trimmed mean awards in
the unexposed control condition were $0 and $71,943, respectively ($SD = $1,302,768).14
But when the taint of prejudicial PTP was introduced (in the exposed treatment condition), median and 5% trimmed mean awards increased to $50,000 and $345,178, respectively ($SD = $710,960). This substantial effect is statistically significant, $U = 4,505$, $Z = 3.127$, $p = .002$, $r = .24$. Even when defense verdicts are excluded, the median damages award increased, more than tripling, from $150,000 to $500,000, $p = .02$.15

**The accuracy of jurors’ self-diagnoses**

If, as *Skilling* prescribes, jurors who admit bias or are unsure of bias are removed, do the effects of pretrial publicity disappear? No, the *Skilling* protocol failed; very few people admitted bias in the first place, and those who did were equally likely to impose liability. Referencing Table 2, the vast majority of jurors denied bias and instead expressed a sureness that they would be able to impartially consider only the evidence presented at trial (91% and 87% in the unexposed control and exposed treatment conditions, respectively). Notably, comparing those who deny bias against those who are either unsure or admit bias across conditions, we find no significant difference, $\chi^2 (1)$

---

14 Given the high variability of damages awards, particularly the presence of a handful of extreme outliers, the mean can be misleading. Here, for example, the mean for the control condition ($292,540) is over $220,000 larger than the trimmed mean of $71,942, and this difference is driven by a single outlying score of $10,000,000.

15 The biasing effect of pretrial publicity was not as clearly reflected in certainty scores (not shown), a measure that is not asked of real jurors. The mean rating, from 1 (“certainly not medical negligence”) to 6 (“certainly medical negligence”) was 3.23 in the control condition, and only increased to 3.61 in the treatment condition—a trending but non-significant result, $t(172) = -1.581$, $p = .116$. Although one might assume that binary verdicts can be directly inferred by splitting the scale in half (i.e., those finding for the defendant rate the case from 1-3, while those finding against the defendant provide a 4-6 rating), that is apparently not the case for all subjects. Just under 5% (4 of 94) of those ruling in favor of the defendant nonetheless gave a rating of 4 or higher; a similar number (3 of 80) of those against the defendant nonetheless gave a rating of 3 or lower. Screening of jurors made little difference on this measure.
= 0.523, \( p = .470 \). That is, subjects were equally likely to admit bias (or not) regardless of whether they read irrelevant or prejudicial pretrial materials. (Again chi-square statistics from the contingency table are reported for ease of explication, but regression analyses reveal equivalent results; see Attachment C).

Excluding pursuant to the *Skilling* protocol, we are left with 59 and 95 subjects in the control and treatment conditions, respectively. As seen in Table 2 (self-screened row), the self-diagnosing protocol completely failed. The verdict rates remained unchanged almost to the digit: before screening, 35% and 52% of control and treatment subjects found against the defendant; after screening, the rates were 34% and 53%. The 34% versus 53% difference continues to be both significant and meaningful, a shift of 19%, 95% CI [3%, 34%], \( \chi^2 \) (1) = 5.152, \( p = .023 \), \( \phi = .18 \). This difference amounts to an odds ratio of 2.17, 95% CI [1.1, 4.3], or in other words, a more than doubling of the odds of a verdict against the defendant, given exposure to prejudicial PTP.

The failure of the *Skilling* protocol is again on dramatic display in the pain and suffering awards. (See Table 2, self-screened row). The median and trimmed mean awards in the control condition remained about the same as they were prior to applying the *Skilling* screen: $0 and $70,536, respectively (SD = $1,366,095). And application of the *Skilling* protocol did not cure the taint of prejudicial PTP in the treatment condition; rather, the median and trimmed mean awards remained higher than in the control condition, at $90,000 and $345,404, respectively (SD = $729,406). This difference across conditions continues to be statistically significant, \( U = 3,553, Z = 2.930, p = .003, r = .24 \).
Even when defense verdicts are excluded, median damages awards quintupled, from $100,000 to $500,000, $p = .03$.

**Discussion**

*Does juror self-diagnosis help at all?*

So, far we have compared the exposed treatment group to the unexposed control group, before and after imposing a screen on the basis of juror self-diagnosis.

For the reasons explained in analytic strategy section above (see p. 71), we have defined success as the screened exposed group performing similarly to the screened unexposed group, as this would show that self-diagnoses screens create an “impartial” jury, as the Constitution guarantees. However, even if self-diagnosis does not achieve the gold standard of impartiality, it might still be better than nothing. For that purpose, one might simply examine the exposed condition before and after screening.

Our data fails to support the hypothesis that juror self-diagnosis helps at all, but we cannot rule out competing hypotheses that it helps somewhat. Our best estimate is that self-diagnoses make matters slightly worse. As shown in Table 2, examining the verdict rates within the exposed conditions for all versus self-screened jury pools, the observed verdict rates against the defendant *increase* from 52% to 53%. This 1% difference is far from statistical significance, of course ($\chi^2 (1) = 0.034, p = .854$). Also, examining damages, nonzero median damages remain unchanged after screening.

The confidence interval around our calculated odds ratios provides further elucidation on this point. For all jurors, pretrial publicity doubles the odds of judgment against the defendant (odds ratio = 2.00). For self-screened jurors, the self-diagnoses
actually made matters worse (odds ratio = 2.17). The point estimate 2.17 is the best estimate of the actual odds ratio.\textsuperscript{16} Nonetheless, the confidence interval around our point estimate ranges from 1.1 to 4.3, meaning that while (at traditional levels of statistical significance) we can rule out the hypothesis that self-screening completely eliminates bias (a 1.0 odds ratio), we cannot rule out the hypothesis that self-diagnosis reduces the bias, say by half (a 1.5 odds ratio).\textsuperscript{17} Nor can we rule out the hypothesis that self-diagnosis makes the bias problem twice as worse (a 4.0 odds ratio). It is, of course, difficult to prove a negative. Rather than scholars trying to endlessly falsify smaller and smaller hypotheses about the efficacy of this self-diagnosis procedure, it may be more sensible for those who propose to use this sort of method to instead show that it is at all reliable.\textsuperscript{18} To be sure, our data do not support the hypothesis that asking jurors whether they can be impartial is at all diagnostic of partiality versus impartiality, and our best estimate is that it makes matters slightly worse.

**Opinion-screening as an alternative to the Skilling protocol**

Recall that in *Skilling*, the Supreme Court again endorsed the protocol of judges first asking jurors whether they have any opinion about the case or the parties, then a

\textsuperscript{16} “As one considers values farther and farther from the point estimate, they become less and less consistent with the observed [odds ratio]. By the time one crosses the upper or lower boundaries of the 95% CI, the values are extremely unlikely to represent the true [odds ratio], given the point estimate (that is, the observed [odds ratio]”) (Guyatt et al., 1994, p. 59).

\textsuperscript{17} Indeed, to rule out the hypothesis that self-diagnosis screening reduces bias in half (a 1.5 odds ratio), it would be necessary to quadruple our sample size in the self-screened condition (i.e., from $N = 154$ to $N = 616$) in order to reach a confidence interval of (1.54-3.03), As the competing hypothesis becomes smaller and smaller, the needed sample size to falsify the hypothesis approaches infinity.

\textsuperscript{18} “[T]he proponent of the evidence has the burden of proving that a scientific principle or technique is reliable” (Friedland, Bergman, & Taslitz, 2000, p. 375).
pair of questions about whether jurors can overcome any opinions and thus base a verdict on only the evidence at trial. The latter two questions then become the basis for screening under the *Skilling* protocol; this is also the analytic approach we applied above.

The juror’s task of introspecting to identify an existing opinion (the first question) is a different cognitive task from assessing one’s ability to overcome that opinion (the second and third questions). In terms of the Wilson and Brekke model discussed earlier (see p.71), the opinion question is akin to the first step (awareness of mental contamination), while the questions about suppressing that opinion also implicates steps two (motivation to correct), three (knowledge of direction and magnitude of bias), and four (ability to adjust).

A different reason that the answers to the latter two questions might be unreliable relates to “social desirability” bias (Tourangeau & Yan, 2007, p. 860). A juror, despite accurately self-diagnosing that he or she cannot overcome a lingering bias, might nonetheless publicly insist that he or she will act impartially. The latter questions essentially ask whether a juror will adhere to the social norms of being a good and fair person—whether he or she will uphold a civic responsibility as any decent person would.

Broeder (1964), after observing a series of voir sessions in a federal district court, and interviewing the venirepersons, put it this way:

> On its face, probably no question commonly asked on voir dire seems as innocuous as, "Can you be fair and impartial?" And it was—when addressed to the panel generally. Where particular jurors only were addressed, however, the situation was quite different. The jurors
particularlly addressed were almost uniformly resentful and felt that their integrity had been brought into question (1964, p. 92).

There is pressure to respond consistently with the norm, lest one appear to be a social deviant, one with the undesirable trait of being uncontrollably swayed by bias.

These considerations imply a modification to the Supreme Court’s questioning protocol that might achieve greater diagnosticity. If courts were to exclude anyone who admits or is unsure of even having an opinion about the case, the response might be more useful than the *Skilling* protocol for constructing the fair and impartial jury required by the Constitution.

To assess the possible efficacy of such a revised protocol, we compared the verdict rates of the pre-*Skilling* protocol jury pool to a jury pool from which we excluded anyone who admitted (or was unsure about) having an opinion about the defendant. Out of 95 jurors that insisted that they could be fair and impartial in the treatment condition, 61 had admitted (or were unsure about) having an opinion about the defendant, but believed that they could suppress that opinion and focus solely on the evidence at trial. That left 34 who stated they had no opinion whatsoever about the defendant, and thus were eligible under our proposed opinion-eligibility standard.

Despite such an aggressive exclusion protocol, juror bias remains rampant. Prior to the *Skilling* screen, 35% and 52% of control and treatment subjects, respectively, found against the defendant, a bias gap of 17%. After the more stringent screen based on mere opinion, the rates were 33% and 56%, a bias gap of 23%. This difference across control and treatment conditions continues to be both significant and meaningful, $\chi^2 (1)$
3. Bias blind spot during voir dire

\[ z = 4.452, p = .035, \phi = .21; \text{ odds ratio} = 2.53. \] Indeed the bias gap got worse, not better.

The failure of the aggressive opinion-screening protocol is again reflected in the damages awards. The trimmed mean awards in the control condition remained about the same as the bias-screened jurors at \$75,898 (SD = \$1,389,038) versus \$442,320 (SD = \$781,066) in the exposed condition. The six-fold disparity in damages caused by exposure to negative PTP remained, regardless of the more aggressive screening protocol.

**Unexplained behavior of fair jurors disqualifying themselves**

If the juror’s actual bias is not what causes her to say she is biased, then what does cause some jurors to answer the *Skilling* questions differently than other jurors? We hypothesized that perhaps the most earnest and thoughtful jurors may be more likely to admit concerns about their own impartiality. Need for Cognition (NFC) and Cognitive Reflection Test (CRT) scales were administered to explore individual differences in the ability to self-diagnosis bias (see Cacioppo, Petty, & Chuan Feng Kao, 1984, regarding NFC; Frederick, 2005, regarding CRT).

NFC is a personality variable reflecting the extent to which people engage in and enjoy exerting cognitive effort. We hypothesized that those higher in Need for Cognition would be more motivated to fully consider the Supreme Court’s prescreening questions, and thus be more likely to self-diagnose bias caused by exposure to the pretrial article about the defendant.

The CRT is designed to assess the degree to which individuals suppress an intuitive and spontaneous wrong answer in favor of a reflective and deliberative right
answer. We hypothesized that those performing more highly on the Cognitive Reflection Test would be more likely to stop and think carefully about the *Skilling* prescreening questions, rather than relying on a quick intuitive assessment of being unbiased, and thus be more likely to self-diagnose bias caused by exposure to the pretrial article about the defendant.

To test the above possibilities, we fitted a binary logistic regression model predicting *Skilling* eligibility from NFC, CRT, and demographic variables on data from the treatment condition.19 As shown in Attachment C, the model itself was far from significant ($\chi^2(9) = 3.904, p = .918$), and contrary to our hypotheses, neither NFC nor CRT nor any of the demographic variables predicted which jurors would disqualify themselves as *Skilling* ineligible. The direction of effects, however, were at least in the right direction, with higher CRT and higher NFC both associated with higher likelihood a positive self-diagnosis. (The CRT odds ratio = 1.33, 95% CI [0.9, 2.0]; the NFC odds ratio = 1.89, 95% CI [0.7, 5.0]. Nonetheless, given the unreliability of these findings, it remains unclear as to what causes some biased jurors to disqualify themselves. All we know is that, alas, it is not the fact of actually being biased.

19 The control condition was not included because, without knowledge of what biased opinions a juror might or might not harbor, it is impossible to predict the effect of NFC and CRT. In the treatment condition, in contrast, we assume jurors harbor the pretrial publicity bias that we experimentally induced, and thus can articulate hypotheses regarding the ability to self-diagnose that bias.
Part I: General discussion

It is notable that our two experiments involved very different populations, who had very different base rates for imposing liability, and very different levels of confidence in their own ability to be fair and impartial (compare Table 1 and Table 2). In the control condition 13% of the law students imposed liability, while 35% of the respondents in the national sample imposed liability. We also saw very different rates of self-diagnoses, with 45% of law students saying that they were unable to be fair and impartial (or unsure), versus only 9% of the respondents in the online national sample.

Notwithstanding these differences, we saw similar effects of pretrial publicity and a similar failure of the screening protocol to remove that bias. This finding should enhance readers’ confidence in the external validity of our studies; they do not seem to be driven by peculiarities about a particular subject pool.

displays the combined data from the 248 subjects in Experiments 1 and 2, using a dot plot to show verdicts, damages awarded, and whether the juror was screened based on their self-diagnoses as to whether they could be fair and impartial (the Supreme Court’s “Skilling protocol”, as we have called it). This graphic depicts the upwards skew of awards in the exposed condition, compared to the unexposed condition. For the Skilling protocol to successfully remove bias, it would need to edit the distribution on the right to make it appear like the distribution on the left. However, the paucity and improper distribution of juror self-diagnoses shows the failure of this protocol to correct for the induced bias.
Examining verdict rates across both the convenience and national samples, 27 of 95 (28%) unexposed persons found negligence, while 71 of 152 (47%) exposed persons found negligence. This was a substantial difference (odds ratio = 2.18, \( \chi^2 (1) = 7.931, p = .005 \)), and indicates successful induction of a pretrial publicity bias as predicted. The Supreme Court’s method of excluding jurors—those who thought themselves to be unable to be fair or impartial or unsure—removed 46 of our jurors (about 19% of our sample). Nonetheless the method altogether failed to correct this doubling of the odds of a liability verdict, leaving the percentage finding negligence almost unchanged after screening (28% in the exposed condition versus 48% in the exposed condition). The difference remained substantial, and if anything increased (odds ratio = 2.40, \( \chi^2 (1) = 8.332, p = .004 \)).

Examining monetary awards (including $0 verdicts), the 95 unexposed control persons had a median and mean (SD) award of $0 and $95,374 ($304,077). The 152 exposed persons had a median and mean (SD) award of $0 and $275,739 ($460,700). This constitutes a tripling of mean damages awards across conditions, \( U = 8,872, Z = 3.331, p = .001 \). Screening out the jurors based on their self-diagnoses failed to cure this bias, \( U = 6,117, Z = 3.290, p = .001 \).

Examining monetary awards conditional on a finding of negligence, the 27 unexposed persons who found negligence had a median and mean (SD) award of $60,000 and $339,092 ($498,817). The 71 exposed persons who found negligence had a median and mean (SD) award of $500,000 and $594,197 ($518,597). With median damages awards eight times higher in the exposed condition, this difference across
conditions is highly significant, both substantively and statistically, $U = 1,324, Z = 2.920, p = .004$, and indicates that pretrial publicity has an effect above and beyond the impact on verdicts. The application of a self-diagnosis screen again failed to cure, with median damages awards remaining unchanged at $60,000 in the unexposed condition versus $500,000 in the exposed condition, $U = 897, Z = 2.581, p = .01$. 
Figure 3. Plot of combined data from Experiments 1 and 2 by condition, with verdict, damages (capped at $1.5M), and self-screen. Each dot represents a single juror (N = 248), with defense verdicts shown in a block on the bottom and plaintiffs’ verdicts shown by amount of damages awarded, as a dot-histogram with $50,000 bins. Jurors who said they were unable to be fair and impartial or were unsure (the “self-screened”) are shown as yellow-striped. Removal of those jurors does not cure the upwards skew (bias) in the condition where subjects were “exposed” to pretrial publicity adverse to the defendant.
Part II: Self versus other diagnoses

Why does the *Skilling* method of diagnosing juror bias fail? As discussed in the introduction, successfully eradicating “mental contamination” is a difficult cognitive task (see p. 62). This self-diagnosis task requires a juror to engage in counter-factual reasoning to answer this question: Will the exposure cause you to change your decision from what it would have been had you not been exposed to the biasing factor? Analytically, that is a lot to ask. Indeed, one might wonder whether jurors are even *doing* that predictive-counterfactual-comparison task, or if they are instead reporting an aspiration to fulfill their civic duty to serve and serve fairly.

Interestingly, the law takes a different approach for potentially biased judges than it does for potentially biased jurors. Federal law requires that judges disqualify themselves not only for self-perceived bias, but also if the circumstances are such that their "impartiality might reasonably be questioned" (28 USC §455; see e.g., *Caperton v. AT Massey Coal Co., Inc.*, 2009, see also *Merritt v. Hunter*, 1978, explaining that “[e]ven though a judge personally believes himself to be unprejudiced, unbiased and impartial, he should nevertheless certify his disqualification where there are circumstances of such a nature to cause doubt as to his partiality, bias or prejudice”). In principle, this would seem to be an objective and more stringent standard that would require disqualification

---

20 Our experiments suggest that it maybe be unwise for judges to make the decision about whether to disqualify themselves. It may be more sensible to allow a different judge to resolve the threshold question, based on an objective review of the circumstances (i.e., potentially biasing factors). Substantively however, putting aside who makes the decision, the judge-disqualification question is different than the one posed for jurors.
more often. As the Supreme Court explained in Caperton, with language notably in contrast to its reasoning in Skilling:

The difficulties of inquiring into actual bias, and the fact that the inquiry is often a private one, simply underscore the need for objective rules. Otherwise there may be no adequate protection against a judge who simply misreads or misapprehends the real motives at work in deciding the case. The judge’s own inquiry into actual bias, then, is not one that the law can easily superintend or review (2009, p. 883).

Chief Justice Marshall, as we saw before, recognized this difficulty in the jury context, writing that the juror “may declare that he feels no prejudice in the case; and yet the law cautiously incapacitates him from serving on the jury because it suspects prejudice, because in general persons in a similar situation would feel prejudice” (United States v. Burr, 1807, p. 50). Our experiments show that Chief Justice Marshall was correct in his skepticism about juries, just as the Caperton court was skeptical about judges hundreds of years later.

The bias blind spot

It turns out that questioning from the perspective of other people might be a more workable inquiry in psychological terms too. The question essentially shifts from an assessment of, “Am I biased?” to one of, “Would a person in this situation be biased?” The former entails self-diagnosis, while the latter entails other-diagnosis. Research on the “bias blind spot” indicates that, in general, people are better at diagnosing bias in others than in themselves (see Ehrlinger, Gilovich, & Ross, 2005, for a review). Consider a rather straightforward test by Pronin, Gilovich, and Ross (2004). They gave people short descriptions of a host of different biases that have been
documented in the scientific literature (e.g., self-serving bias, halo effect, fundamental attribution error), and simply asked the subjects to rate the extent to which they would succumb to each bias relative to the extent to which the average American would succumb to the same bias. The results were clear: almost without exception, subjects reported that those other persons would be much more susceptible to each of the biases than they themselves would be. This sort of bias blind spot seems to be caused by (at least) three related psychological mechanisms.

The first is motivated reasoning, in particular its manifestation as a self-serving bias effect. Substantial research indicates that people are motivated to perceive themselves in a positive light, including a freedom from unwanted bias or prejudice. Numerous studies have shown, for example, that the large majority of persons consider themselves better than average, despite that being a mathematical impossibility (see e.g., Taylor & Brown, 1988). This effect causes people to overestimate their personal immunity from bias, an overestimation that does not occur in their estimates of bias in others.

The second reason is non-motivational, and relates to the cognitive processing underlying the detection of bias. In particular, there is an asymmetry in the type of information available for self-diagnosis and other-diagnosis (Ehrlinger et al., 2005). With the former, the person is able to introspect, that is, self-reflect on the contents of his or her own mind. With other-diagnosis, in contrast, the only available evidence is the observable behavior of the actor. Unable to peer into the contents of the mind, observers must resort to using the overt behavior, typically coupled with lay theories of
bias (e.g., “money corrupts”). For their own self-assessments, on the other hand, such lay theories will be eschewed in favor of more direct evidence, in particular, introspection. The problem with this tactic is connected to the third reason.

The third reason is illusory introspection. What is meant by this is that introspection for bias is likely to generate false negatives (see Pronin et al., 2004, for a review). Most cognitive processes are unconscious, and thus leave no phenomenological trace within the mind (Bargh & Chartrand, 1999; Bargh & Morsella, 2008). The implication is that there is therefore no phenomenological trace—no consciously accessible evidence—for introspection to possibly detect, even if there is in fact a bias, even a severe bias. In fact, illusory introspection might actual worsen the accuracy of self-diagnosis, since the effort of self-reflection, plus a negative finding, could create a false sense of confidence in one’s objectivity (see e.g., Uhlmann & Cohen, 2007).

This suggests a simple change might improve the accuracy of voir dire: stop asking jurors to self-diagnose bias, and instead ask them to diagnose how the pretrial publicity they have seen would bias others.

**Experiment 3: Self vs. other (undergraduates)**

We test here the hypothesis that jurors will more readily concede that the biasing pretrial publicity they have seen would bias others than that it would bias themselves. If this is the case, then other-diagnosis, unlike self-diagnosis, might provide a viable voir dire tool.
Methods

Participants

Undergraduate students from a large midwestern university were recruited (in exchange for course credit) during the spring 2013 semester. Although 174 subjects enrolled in the study, data cleaning removed the following: 33 cases because incomplete (i.e., subject exited study before completion); 2 persons for finishing in less than 32 minutes (which is impossible, since the video itself lasts 32 minutes); 1 person who was less than 18 years old (and thus jury ineligible); and 11 cases for missing 2 or more of 3 quality assurance questions. The requirement to exclude 27% of the data for non-compliance decreased the effective sample size much more than anticipated.

The final sample (N = 127) was predominantly white (75%), tended to be male (54%), and had an average age of about 20. Almost everyone indicated—not surprisingly, given how they were recruited—that they had completed high school and

21 “Quality assurance” questions are those with directed answers, such as, “Unless you are randomly clicking, please select ‘Strongly Agree.’” A participant’s failure to select the directed answer thus provides evidence of noncompliance with the task. The accuracy rate for each of the four quality assurance questions was in the 86%-90% range. However, it seems that even genuine subjects would, presumably due to an occasional lapse of attention, miss at least one such question: although 77% of participants answered all 3 correctly, a substantial minority (15%) missed at least one question. 8%, however, missed 2 or more. The goal of screening was to remove cheaters, not screen for high attention, and thus we opted to select an error rate of 2 or more as the threshold for removal.

22 The complete distribution was as follows: 75% White; 3% Black or African American; 2% American Indian or Alaskan Native; 8% Asian; 0.1% Native Hawaiian or Other Pacific Islander; and 12% “some other race.”

23 The mean, median, and standard deviation of age were, respectively, 19, 19, and 1.2.
were thus in college. Thus the sample was skewed toward white, educated, younger males. However, these demographic characteristics were successfully distributed at random throughout the experimental conditions.

**Procedure**

The same stimuli and analytic approach as in Experiments 1 and 2 were used. The only difference is that we added an extra experimental variable, namely, whether the diagnostic task was framed in the first or third person. In particular, we use a 2×2 between-subjects experimental design. Subjects were, as before, randomized to read either an irrelevant article or an article casting the defendant in a highly negative light. All subjects then watched the same 32-minute medical malpractice trial video. As part of the mock trial, there was a judge-juror colloquy during voir dire wherein the subject was asked either whether *he or she* was biased (“first person” voir dire) or—here is where the method differs from the earlier studies—whether *other jurors*, if exposed to the same circumstances as the subject, would be biased (“third person” voir dire). See above at p. 77 for the transcript of the first person voir dire; recall the two key questions were:

1. “Would any opinion you have prevent your impartial consideration?

---

The complete distribution, of highest obtained educational level, was as follows: 0.0% Elementary School; 0.8% Some High School; 44% High School Graduate; 54% Some College Credit; 0.8% Associate Degree; 0.8% Bachelor’s Degree; 0% Master’s Degree; 0% Professional Degree; and 0% Doctoral Degree.

That 44% said they were high school graduates rather than had completed some college credit is baffling. Some might actually be in the first semester of college (and thus actually not have “completed” any college credit yet). However, this study was carried out in the spring term, and thus it is unlikely that almost half of the sample were first semester students. This would suggest students were confused on the question. On the other hand, it is possible that the introductory psychology courses from which the sample was drawn are skewed toward entering students.
of the evidence at trial?”; and

2. “Could you base a verdict only on the evidence at trial?”

In the third person condition, everything was the same, except the two key diagnostic questions were framed in the third person, as follows:

1. “I understand that you will do your best to be fair and impartial, but do you think that there is a significant risk that other jurors, exposed to the article you read, would be prevented from impartially considering the evidence at trial?”; and

2. “Do you think those other jurors, exposed to the article you read, would base their verdict only on the evidence at trial?”

**Results**

**The effect of pretrial publicity**

Table 4 shows the percentage of subjects finding for the plaintiff, broken down by voir dire question (1st vs. 3rd person), type of article (control vs. the treatment negative PTP), and pre and post screening (i.e., pre and post removal of subjects, per the *Skilling*-based analytic approach). As can be seen, subjects exposed to the biasing pretrial publicity were, prior to any screening for bias, more likely to find for the plaintiff, albeit not at traditional levels of statistical significance. In the first person condition, the rates increased from 52% to 62%, a bias gap of 10%, 95% CI [-13%, 35%], \( p = .40 \). In the third person condition, the rates increased from 34% to 53%, a bias gap of 19%, 95% CI [-4%, 42%], \( p = .13 \). Unfortunately, the failure to cleanly induce a strong bias undermines our ability to assess diagnostic accuracy—there must be a bias actually present to detect. Nonetheless, the confidence intervals indicate bias is more likely than not; given this fact, and our results from experiments 1 and 2 (plus the larger literature on PTP)
showing the biasing potential of this stimuli, we can reasonably press forward and assume that some degree of bias has been induced.

Figure 4. Verdict rates by voir dire, pretrial publicity, and screen, for the UA undergraduate pilot sample in Experiment 3. The 1st person voir dire screen failed to remove the biasing effect of pretrial publicity, and if anything worsened the bias gap. The 3rd person voir dire screen removed most of the subjects, thereby rendering the post-screen data highly unstable (see text).

**First versus third person diagnoses**

Table 4 show diagnoses across conditions. Looking first at the 1st person voir dire block, we see that almost half of subjects in the negative PTP (exposed) condition conceded they were unsure or outright biased. Eighteen out of 32 (56%) insisted they were impartial. This is a lower rate than observed in the irrelevant article (unexposed) condition, where 21 out of 29 (72%) confidently asserted impartiality, although the difference is only a handful of people and not reliable ($p = .19$). Regardless the important question is what happens to verdicts, which we turn to in a moment. But first also look at the 3rd person results.
Table 4. Rates of diagnosing bias by voir dire and pretrial publicity, for the undergraduate sample. Subjects were, as hypothesized, significantly more likely to diagnose bias in others than in themselves.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Insist Impartial</th>
<th>Unsure</th>
<th>Admit Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st}</td>
<td>Exposed</td>
<td>32</td>
<td>18 (56%)</td>
<td>6 (19%)</td>
</tr>
<tr>
<td></td>
<td>Unexposed</td>
<td>29</td>
<td>21 (72%)</td>
<td>3 (10%)</td>
</tr>
<tr>
<td>3\textsuperscript{rd}</td>
<td>Exposed</td>
<td>34</td>
<td>3 (9%)</td>
<td>24 (71%)</td>
</tr>
<tr>
<td></td>
<td>Unexposed</td>
<td>32</td>
<td>5 (16%)</td>
<td>15 (47%)</td>
</tr>
<tr>
<td>n</td>
<td>N = 127</td>
<td>47</td>
<td>48</td>
<td>32</td>
</tr>
</tbody>
</table>

In the 3\textsuperscript{rd} person voir dire block, participants were, as hypothesized, significantly more willing (than in the 1\textsuperscript{st} person voir dire) to diagnose the corrupting potential of the negative PTP. Only 3 out of 34 (9\%) of those exposed to the biasing pretrial publicity were confident that another person would not be tainted by the information, which is dramatically different than that in the self-diagnosis condition. An unexpected result, however, was that subjects were also more likely to attribute possible bias to another person even in the unexposed condition. Only 5 out of 32 (16\%) insisted on impartiality in this condition; this, too, is substantially different than observed for the 1\textsuperscript{st} voir dire, although it is a smaller effect. This suggests that subjects might diagnose bias more in others, but not accurately—which would undermine the efficacy of the 3\textsuperscript{rd} person voir dire.

\textit{First versus third person diagnostic accuracy}

As with Experiments 1 and 2, we once again find no evidence that jurors can successfully self-diagnosis their bias. Post-screen, in the 1\textsuperscript{st} person condition (see Figure
3. Bias blind spot during voir dire

), we actually observe the bias gap worsen. Forty-three percent and 67% of jurors in the control and treatment groups, respectively, now find against the defendant—an absolute difference of 24%, 95% CI [-7%, 54%], \( p = .14 \).

It is unfortunately impossible to interpret diagnostic accuracy in the other-person condition. The problem is that too many people were screened out in both conditions, such that only eight jurors remain—five in the unexposed control condition, and three in the exposed condition. The verdict rate actually flip-flops to 60% in the unexposed group and 30% in the treatment group, but this 30% “difference” is completely uninformative given the sample size, \( p = 1.0 \).

**Discussion**

The first and third person voir dire protocols reveal clear evidence of a bias blind spot. Subjects were very willing to concede that *another person* would be biased by the negative PTP they had read, while simultaneously insisting that they themselves, in contrast, would not be so affected. This suggests that voir dire might be improved if the impartiality screening were framed in 3rd rather than 1st person terms.

However, subjects were more likely to diagnose biases in others regardless of whether the article in question was neutral or biasing. This suggests the 3rd person voir dire might result in more screening, but not more *accurate* screening. This is an undesirable result, since it would remove venirepersons—an expense, since the court would need to find replacements—without any concomitant benefit (viz., a more impartial jury panel). Unfortunately, the pilot sample turned out to be too small to fully assess this issue of accuracy. So many subjects removed themselves in the 3rd person
voir dire that meaningful statistical analyses were impossible. Experiment four tries again, using a larger and more diverse sample.

**Experiment 4: Self vs. other (MTurk)**

**Method**

**Participants**

A sample of 673 persons were recruited via Amazon Mechanical Turk, for $3 payment, between May 10th and 15th, 2013. Four steps were taken to screen for quality data: (1) 31 cases were removed for impossible completion times (viz., finishing in less than 32 minutes, which is the length of the trial video); (2) 1 case was removed wherein the subject was less than 18 years old and thus jury-ineligible; (3) written entries were examined for evidence of non-response (e.g., copy-and-pasting the question prompt, or smashing random key strokes), although no cases were identified as suspicious; and (4) 17 cases were removed for missing 2 or more of 4 quality assurance questions.25 Interestingly, whereas we had to remove 27% of the undergraduate data as junk in Experiment 3, here we only had to remove 7%.

---

25 Although 87% of participants answered all 4 correctly, a substantial minority (11%) missed at least one question. Less than 2%, however, missed 2 or more. The goal of screening was to remove cheaters, not screen for high attention, and thus we opted to select an error rate of 2 or more as the threshold for removal.
The final sample \((N = 624)\) was predominantly white (80%), tended to be female (57%), and had an average age in the 30s. Most (78%) had taken at least some college coursework, and a notable minority (10%) had additional education at the Master’s or higher level. Thus the sample was skewed toward white, educated, relatively younger females, a not uncommon finding for Amazon Mechanical Turk samples. However, these demographic characteristics were successfully distributed at random throughout the experimental conditions.

**Procedure**

The same procedure as reported in Experiment 3 was used.

**Results**

**The effect of pretrial publicity**

Figure 5 shows the percentage of subjects finding for the plaintiff split by voir dire question (1st vs. 3rd person), type of article (control vs. the treatment negative PTP), and pre and post screening (i.e., pre and post removal of subjects, per the *Skilling*-based analytic approach). As can be seen, subjects exposed to the negative PTP were, prior to any screening or bias, more likely to find for the plaintiff: 47% vs. 41% in the 1st person voir dire, and 48% vs. 42% in the 3rd person voir dire. Similar to Experiment 3, however,

26 The complete distribution was as follows: 80% White; 10% Black or African American; 2% American Indian or Alaskan Native; 6% Asian; 0.1% Native Hawaiian or Other Pacific Islander; and 2% “some other race.”

27 The mean, median, and standard deviation of age were, respectively, 35, 32, and 12.

28 The complete distribution, of highest obtained educational level, was as follows: 0.0% Elementary School; 0.6% Some High School; 11% High School Graduate; 31% Some College Credit; 13% Associate Degree; 35% Bachelor’s Degree; 8% Master’s Degree; 1% Professional Degree; and 0.8% Doctoral Degree.
we failed to induce a PTP bias as strongly as hoped (and as observed in the first two experiments), although the trends are as expected. Exposure to the treatment PTP caused a 6-percentage point increase in the 1\textsuperscript{st} person condition, 95\% CI [-5\%, 17\%], \( p = .34 \), and a similar 6-percentage point increase in the 3\textsuperscript{rd} person condition, 95\% CI [-5\%, 17\%], \( p = .34 \).

**Figure 5.** Verdict rates by voir dire, pretrial publicity, and screen, for the online mTurk sample in Experiment 4. The 1\textsuperscript{st} person voir dire screen failed to remove the biasing effect of pretrial publicity, and if anything worsened the bias gap. The 3\textsuperscript{rd} person voir dire screen removed many of the subjects, and seems to have lessened the bias gap, but the analyses remain underpowered.

**First versus third person diagnoses**

Rates of diagnoses are shown in Table 5. We replicate for a fourth time that very few people self-diagnose bias in the 1\textsuperscript{st} person voir dire condition. 128 out of 158 (81\%) of those exposed to biasing PTP nonetheless insisted, without doubt, that (1) no opinion would prevent their impartial consideration of the evidence at trial and that (2) they
could successfully base their verdict only on the evidence at trial. This is a similar rate as observed in the neutral condition, where 133 out of 155 (86%) confidently asserted impartiality.

Table 5. Rates of diagnosing bias by voir dire and pretrial publicity, for the undergraduate sample. Subjects were, as hypothesized, significantly more likely to diagnose bias in others than in themselves.

<table>
<thead>
<tr>
<th></th>
<th>1st Exposed</th>
<th></th>
<th></th>
<th>1st Unexposed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Insist Impartial</td>
<td>Unsure</td>
<td>Admit Bias</td>
<td>Insist Impartial</td>
<td>Unsure</td>
</tr>
<tr>
<td>1st Exposed</td>
<td>158</td>
<td>128 (81%)</td>
<td>22 (14%)</td>
<td>8 (5%)</td>
<td>133 (86%)</td>
<td>15 (10%)</td>
</tr>
<tr>
<td>1st Unexposed</td>
<td>155</td>
<td>133 (86%)</td>
<td>15 (10%)</td>
<td>7 (5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Exposed</td>
<td>155</td>
<td>38 (25%)</td>
<td>42 (27%)</td>
<td>75 (48%)</td>
<td>56 (36%)</td>
<td>58 (37%)</td>
</tr>
<tr>
<td>3rd Unexposed</td>
<td>156</td>
<td></td>
<td></td>
<td></td>
<td>56 (36%)</td>
<td>58 (37%)</td>
</tr>
</tbody>
</table>

In the 3rd person voir dire condition, in contrast, participants were significantly more willing to diagnose the contaminating potential of the negative PTP. Only 38 out of 155 (25%) of those exposed to the biasing PTP were confident that another person would not be tainted by the information, which is substantially different than that observed in the self-diagnosis condition. In particular, subjects exposed to biasing pretrial publicity were more than three times as likely to diagnose bias in another person than to diagnose possible bias in themselves (OR = 3.30, 95% CI [2.16, 5.05], \( \chi^2(1, N = 479) = 31.97, p < .0001 \)). Subjects were also likely to attribute possible bias to another person even in the neutral unexposed condition. Only 56 out of 156 (36%) insisted on impartiality in this condition; this, too, is significantly different than observed for the
self-diagnosis voir dire, although it is a smaller effect (OR = 2.39, 95% CI [1.63, 3.51], \( \chi^2(1, N = 500) = 20.29, p = < .0001. \)

**First versus third person diagnostic accuracy**

Our analytic approach was yet again undermined by the fact that the negative PTP stimuli failed to significantly bias subjects. Nonetheless the pattern of results fits our hypotheses. In particular, in the first person voir dire condition, the biasing effect of the negative PTP persists; indeed, the bias gap actually worsens (see 4). Prior to the screen, there was a 6% bias gap, 95% CI [-5%, 17%], \( p = .34; \) after the screen, it was 52% versus 41%, an 11% bias gap, 95% CI [-1%, 23%], \( p = .08. \)

The third person voir dire screen is again, unfortunately, nearly impossible to analyze with any statistical confidence. But the pattern was at least not in the wrong direction. The pre-screen bias gap was 6%, 95% CI [-5%, 17%], \( p = .34; \) after the screen, it shrinks to 45% vs. 41%, or only a 4% bias gap, 95% CI [-17%, 24%]. Again though, trends should be interpreted with caution.

**Cognitive reflection and diagnoses**

A preliminary analysis reveals that differences in cognitive traits, notable the Need for Cognition, might provide a window into why bias diagnoses differences depending on first versus third person perspective. Two logistic regression models were constructed, one for each voir dire condition, wherein diagnosis of bias (1 = unsure/admit bias; 0 = insist impartial) was regressed on CRT score (range 0-3, with higher scores indicating higher “need for cognition”). Results indicated that CRT score was not predictive of other-diagnosis (\( p = .27 \)), but—notably—that higher CRT scores
were correlated with less willingness to self-diagnose bias \( (p = .06) \). Specifically, the coefficient for CRT was -0.244, indicating that with each additional correct answer on the CRT test the subject’s log odds of self-diagnosing bias decreased by .244 (which amounts to an odds ratio of 0.78).

**Part II: General discussion**

The attempt to assess the relative accuracy of 1\textsuperscript{st} vs. 3\textsuperscript{rd} person voir dire was, unfortunately, stymied by the fact that the negative pretrial publicity stimuli failed to adequately bias the subjects. Recall we were not testing whether pretrial publicity can bias—a wide range of research shows it can—but rather wanted to use it simply to bias subjects, in order to then ask the question of whether they could diagnose the bias. Since the initial biasing failed, we cannot answer the question of interest with much statistical confidence.

Nonetheless, the results continue to be promising in two respects. First, there is a clear difference (statistically supported) in the willingness to diagnose bias: people more readily diagnose bias in others than in themselves. This was shown in both the pilot and MTurk sample. Thus the bias blind spot shows up in important applied settings, in this instance the courtroom.

Second, although the undergraduate sample in Experiment 3 was (due to small sample size) far too unstable to assess 3\textsuperscript{rd} person voir dire accuracy, the MTurk sample in Experiment 4, with a larger size \( (n = 94) \), at least shows a trend of shrinking the bias gap. These two facts suggest that the potential of other framing voir dire should be subjected to additional testing, before it is ruled out. It shows promise.
3. Bias blind spot during voir dire

Limitations

Our study had several limitations. First, we tested a particular type of juror bias—the bias due to pretrial publicity, and indeed used a particular form of pretrial publicity based on a real newspaper article.\(^{29}\) Publicity, unlike many other sources of biases, is amenable to manipulation and thus randomized experimentation. Whether and to what extent jurors are biased from other sources, and whether they may be better able to self-diagnose those biases, are open questions.

Second, we should emphasize that we have not calibrated the degree of bias observed in our mock jurors with the amount of bias that may or may not have infected any particular juror in any particular case. (Nor could we.) It is possible that none of the jurors in the *Skilling* case were biased at all, or that they were collectively more biased than the mock jurors in our study. Relatedly, our data cannot say whether the amount of bias shown in our study (a more than doubling of odds of imposing liability on the defendant) is “too much” bias to be constitutionally tolerable.\(^{30}\) Maybe a court would say that a doubling of odds is “good enough impartiality,” noting that nearly half of the

\(^{29}\) Not unlike real instances of pretrial publicity, our stimulus was complex, consisting of multiple components in a 1300-word article. Thus, we are unable to say which aspects of our stimulus were most biasing. Was it the prior instances of misconduct attributed to the physician-defendant? Was it the discussion of the suffering of another plaintiff? Perhaps the large numbers mentioned in the stimulus, even had an anchoring effect? Future research could try to disaggregate these effects. One might speculate that jurors would be better able to self-diagnose for some causes of biases over others.

\(^{30}\) Presumably, a small amount of partiality is permissible. *See e.g.*, *Skilling*, 130 S.Ct. at 2914 (describing prior cases as ones where the “trial atmosphere was utterly corrupted by press coverage”); *id.*, at 2913 (asking whether there was “extraordinary local prejudice”); *id.*, at 2922 (describing the “deep and bitter prejudice” in a prior case). Notably, however, some of these quotations refer to the distinct legal question of whether prejudice should be presumed, as distinct from the question of whether a particular juror suffered from actual prejudice. *See id.*, at 2917 (making this transition in two different analyses).
3. Bias blind spot during voir dire

jurors still exonerated the defendant even with the publicity. Regardless of such line

drawing problems, our experiment instead focuses on the epistemic device that the
courts use to assess bias. In this sense, our study does suggest that if courts use this
device to reassure themselves that jurors were in fact unbiased—below whatever
threshold for too much bias they select—then the courts’ conclusions are unwarranted
on that basis.

Third, we used a 32-minute abridged civil trial for our experimental stimulus.
The condensed stimulus allowed us to utilize a randomized controlled trial experimental
design, which is the gold-standard for scientific research. Still, there are reasonable
concerns about external validity. Specifically, ours was a civil trial, but pretrial publicity
problems often arise in a criminal context instead. We are unaware of any evidence that
jurors called for criminal trials are somehow better able to diagnose their own biases
than jurors called for civil trials, though one could speculate that the difference in
standards (preponderance versus beyond a reasonable doubt) could matter. The length
of our stimulus also raises the possibility that in a real trial, which may last for days or
even weeks and where biasing factors are less proximate, jurors would more heavily
weigh the evidence presented therein, and thus be less subject to pretrial publicity
biases at all.\footnote{In fact, we observed some heterogeneity even within our sample as to who people react to pretrial
publicity. At least in terms of binary verdicts, many subjects voted the same way as they would have in the
unexposed condition, since the exposure only created a 19% bias gap on the margin in the national
sample. Indeed, we used this same stimulus with another convenience sample of law students and found
an insignificant biasing effect, which thus prevented us from testing whether subjects could self-diagnose
those biases in that sample. The experiments reported herein presume the existence of a biasing factor.}

These limitations apply more to the question of whether jurors are
3. Bias blind spot during voir dire

biased, as opposed to whether they are able to self-diagnose that bias, the question studied here.

Fourth, the experiments were conducted with convenience samples of law and undergraduate students and national samples of human subjects online, who reached individual judgments rather than collective jury verdicts after deliberation. It is possible that real jurors in real courthouses are somehow better able to diagnose their own biases.\textsuperscript{32} Prior research has shown that “the population of Mechanical Turk is at least as representative of the U.S. population as traditional subject pools” (Paolacci, Chandler, & Ipeirotis, 2010). Known experimental results have been replicated using the MTurk population (Berinsky, Huber, & Lenz, 2012). Still, it is likely that MTurkers may be more easily distracted from the trial compared to real jurors, and they may even provide junk responses. Although we paid respondents rather generous bonuses contingent on their measurable performance on attention tasks, such problems could increase noise in the data. It may be that real jurors are more earnest in their efforts to diagnose their own biases. On the other hand, real jurors may have other motivations for saying that they are unbiased (if they have an axe to grind against the defendant, or the social pressure

\textsuperscript{32} But see (Suggs & Sales, 1980) (arguing that courtrooms are particularly bad contexts for elucidating candid responses, given modes of questioning, interaction distance, and formality, which potentially makes real world voir dire practices even worse than the experimental procedures employed here) and (Rose & Diamond, 2008) (reviewing evidence that trial judges browbeat jurors into saying that they can be fair and will do their duty, thus likely reducing the sensitivity of their self-diagnoses). (See also O’Connor, Connolly, Davis, & Sales, 1994, showing that instructing jurors on the law prior to conducting voir dire had no effect on the jurors’ responses).
of answering in public) or biased (if they would prefer not to serve on the jury), which would further reduce the diagnosticity of the questioning.

Fifth, we merely tested whether jurors could diagnose their own biases or predict other’s biases, and thereby provide reliable information to the judge tasked with deciding whether to exclude the juror, or change the venue. One could speculate that “by looking the juror in the eye,” as the Supreme Court suggests, judges are able to ascertain whether he or she can be impartial in a more holistic way (*Skilling v. US*, 2010, p. 2924). Although there are reasons to doubt this would be the case (e.g., Kerr et al., 1990, testing judges who reviewed videotapes of potential jurors questioned about exposure to pretrial publicity, and finding no correlation between the judge’s assessments and the juror’s verdict votes). Attorneys may also use such a holistic approach and their limited number of peremptory challenges to exclude jurors. Our study merely suggests that, in making that holistic assessment, courts and attorneys should give no weight to the content of the juror’s own professions of impartiality or partiality.

**Conclusions**

The courts of appeal say that they will defer to trial court determinations as to whether a juror can be impartial, as long as those determinations are based on “substantial evidence” (see e.g., *California v. Boyette*, 2003, p. 414; see also Hannaford-Agor & Waters, 2004, p. 3, noting that “[a]s a practical matter . . . all U.S. jurisdictions give substantial discretion to trial judges with respect to these decisions”). This study has shown that the juror’s self-diagnoses of bias does not provide substantial evidence
as to their actual impartiality. Indeed it doesn’t seem to provide any evidence: Jurors here completely failed to self-diagnose bias. Trial courts should not rely upon such unreliable answers.

We tried to test a different tactic, namely, rather than asking jurors to make any self-assessments about their own bias, instead ask if others would be biased by the exposure. Our evidence is consistent with a bias blind spot, and suggestive this remedy has potential. But ultimately we lacked power to adequately test the proposition, since the bias induction was too weak and sample size-post screening to small.

In sum, these findings challenge a longstanding and ubiquitous practice of the state and federal courts, used in both civil and criminal trials. Nearly 30 years ago, the Supreme Court said that, “[i]t is fair to assume that the method we have relied on since the beginning . . . usually identifies bias” (Patton v. Yount, 1984, p. 1038, citing Burr). Our study undermines that assumption. Although further research is warranted, it is now fair to put the burden on those who rely upon this particular method of diagnosing bias to show that such reliance is reasonable. One can no longer simply “assume” that it is.
**Attachment B. Regression predicting verdict.** Logistic regression predicting the odds of a plaintiff verdict, with nested comparisons for national sample Data. The overall model regressing verdict on condition was statistically significant for the jury both before (Model A1, $\chi^2 (1) = 4.738, p = .029$) and after (Model A2, $\chi^2 (1) = 5.218, p = .022$) applying the self-diagnosis screen, revealing that self-diagnosis failed to cure the bias induced via pretrial publicity. Nested comparisons revealed that adding self-diagnosis, CRT, and NFC (Model B) failed to improve the model, $\chi^2 (3) = 1.06, p = 0.787$; addition of demographic variables (Model C) likewise failed to improve the model, $\chi^2 (4) = 0.34, p = .987$. Thus the observed effects can be confidently attributed solely to the experimental manipulations.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>B</th>
<th>S.E.</th>
<th>Wald's $\chi^2$</th>
<th>p</th>
<th>Lower</th>
<th>Odds Ratio</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model A1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_treat</td>
<td>0.69</td>
<td>0.32</td>
<td>4.63</td>
<td>.031</td>
<td>1.06</td>
<td>2.00</td>
<td>3.77</td>
</tr>
<tr>
<td>constant</td>
<td>-0.60</td>
<td>0.26</td>
<td>5.39</td>
<td>.020</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Model A2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_treat</td>
<td>0.77</td>
<td>0.34</td>
<td>5.07</td>
<td>.024</td>
<td>1.11</td>
<td>2.17</td>
<td>4.25</td>
</tr>
<tr>
<td>constant</td>
<td>-0.67</td>
<td>0.28</td>
<td>5.90</td>
<td>.015</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Model B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_treat</td>
<td>0.71</td>
<td>0.33</td>
<td>4.80</td>
<td>.028</td>
<td>1.08</td>
<td>2.04</td>
<td>3.86</td>
</tr>
<tr>
<td>self-diagnosis</td>
<td>-0.16</td>
<td>0.49</td>
<td>0.11</td>
<td>.743</td>
<td>0.33</td>
<td>0.85</td>
<td>2.23</td>
</tr>
<tr>
<td>CRT</td>
<td>-0.44</td>
<td>0.13</td>
<td>0.12</td>
<td>.727</td>
<td>0.75</td>
<td>0.96</td>
<td>1.22</td>
</tr>
<tr>
<td>NFC</td>
<td>0.30</td>
<td>0.32</td>
<td>0.93</td>
<td>.335</td>
<td>0.73</td>
<td>1.35</td>
<td>2.51</td>
</tr>
<tr>
<td>constant</td>
<td>-0.56</td>
<td>0.54</td>
<td>1.09</td>
<td>.300</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Model C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition_treat</td>
<td>0.70</td>
<td>0.33</td>
<td>4.60</td>
<td>.032</td>
<td>1.06</td>
<td>2.01</td>
<td>3.81</td>
</tr>
<tr>
<td>gender</td>
<td>-0.12</td>
<td>0.32</td>
<td>0.00</td>
<td>.970</td>
<td>0.53</td>
<td>0.99</td>
<td>1.84</td>
</tr>
<tr>
<td>age</td>
<td>-0.94</td>
<td>0.15</td>
<td>0.06</td>
<td>.801</td>
<td>0.71</td>
<td>0.96</td>
<td>1.30</td>
</tr>
<tr>
<td>race</td>
<td>0.16</td>
<td>0.38</td>
<td>0.19</td>
<td>.664</td>
<td>0.56</td>
<td>1.18</td>
<td>2.49</td>
</tr>
<tr>
<td>education</td>
<td>-0.02</td>
<td>0.15</td>
<td>0.01</td>
<td>.918</td>
<td>0.74</td>
<td>0.99</td>
<td>1.31</td>
</tr>
<tr>
<td>constant</td>
<td>-0.49</td>
<td>0.61</td>
<td>0.66</td>
<td>.417</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Attachment C. Regression predicting self-diagnosis. Logistic regression predicting the odds of a self-diagnosis for national sample data. The overall model regressing self-diagnosis on condition, CRT, and NFC was not statistically significant, $\chi^2 (3) = 4.806, p = .187$, revealing that jurors equally likely (or rather unlikely) to self-diagnosis bias regardless of exposure to pretrial publicity or psychological constructs related to self-reflection and tendency to think.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$B$</th>
<th>S.E.</th>
<th>Wald’s $\chi^2$</th>
<th>$p$</th>
<th>Lower</th>
<th>Odds Ratio</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition_treat</td>
<td>-0.40</td>
<td>0.52</td>
<td>0.58</td>
<td>.45</td>
<td>0.24</td>
<td>0.67</td>
<td>1.88</td>
</tr>
<tr>
<td>CRT</td>
<td>0.29</td>
<td>0.20</td>
<td>2.07</td>
<td>.15</td>
<td>0.90</td>
<td>1.33</td>
<td>1.97</td>
</tr>
<tr>
<td>NFC</td>
<td>0.63</td>
<td>0.50</td>
<td>1.58</td>
<td>.21</td>
<td>0.70</td>
<td>1.89</td>
<td>5.02</td>
</tr>
<tr>
<td>constant</td>
<td>-0.49</td>
<td>0.61</td>
<td>0.66</td>
<td>.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
References


California v. Boyette, 58 P. 3d 391 (Supreme Court 2003).

Caperton v. AT Massey Coal Co., Inc., 556 US 868 (Supreme Court 2009).


Guyatt, G. H., Sackett, D. L., Cook, D. J., Guyatt, G., Bass, E., Brill-Edwards, P., ... others. (1994). Users’ guide to the medical literature: II. How to use an article about therapy or prevention. What were the results and will they help me in caring for my patients? *Jama, 271*(1), 59–63.


Magna Trust Co. v. Illinois Cent. R. Co., 728 NE 2d 797 (Appellate Court, 5th Court 2000).


3. Bias blind spot during voir dire


Patton v. Yount, 467 US 1025 (Supreme Court 1984).


Reynolds v. United States, 98 US 145 (Supreme Court 1879).


Sawyer v. Southwest Airlines Co. (10th Cir. 2005).

Skilling v. US, 561 US 358 (Supreme Court 2010).


State v. Addison, 160 NH 493 (Supreme Court 2010).


3. Bias blind spot during voir dire

*Human Decision Processes, 104*(2), 207–223.


CHAPTER 4: THE EFFECT OF BLINDED EXPERTS ON JUROR VERDICTS

Author Note

Abstract

“Blind expertise” has been proposed as an institutional solution to the problem of bias in expert witness testimony in litigation. At the request of a litigant, an intermediary selects a qualified expert and pays the expert to review a case without knowing which side requested the opinion. This paper reports an experiment that tests the hypothesis that, compared to traditional experts, such “blinded experts” will be more persuasive to jurors. A national sample of mock jurors \(N=275\) watched an online video of a staged medical malpractice trial, including testimony from two medical experts, one of which (or neither, in the control condition) was randomly assigned to be a blind expert. We also manipulated whether the judge provided a special jury instruction explaining the blinding concept.

Descriptively, the data suggest juror reluctance to impose liability. Despite an experimental design that included negligent medical care, only 46 percent of the jurors found negligence in the control condition, which represents the status quo.

Blind experts, testifying on either side, were perceived as significantly more credible, and were more highly persuasive, in that they doubled (or halved) the odds of a favorable verdict, and increased (or decreased) simulated damages awards by over $100,000. The increased damages award appears to be due to jurors hedging their damages awards, which interacted with the blind expert as a driver of certainty. Use of a blind expert may be a rational strategy for litigants, even without judicial intervention in the form of special jury instructions or otherwise.
Background

The American legal system tasks judges and jurors – both laypersons as to science – with resolving highly-technical questions. These laypersons are asked, for example, to evaluate DNA evidence to determine whether it inculpates a particular defendant, to determine the standard of care for lumbar radiculopathy, to interpret epidemiological data to determine whether a given chemical causes an observed disease, and to ascertain the state of the art in a patent suit for computer software. Thus, in both civil and criminal litigation, expert witnesses play an increasingly prominent and important role to inform those layperson decision makers. In civil litigation, experts appear in the vast majority of trials (Gross, 1991), and courts have held in the criminal context that the right to effective counsel includes the right to funding for expert testimony (Ake v. Oklahoma).

In the American legal system, expert witnesses are hand-picked by the litigants, coached towards favorable opinions, and compensated for continuing work only as long as those opinions are favorable (Robertson, 2010). These experts are thus subject to various cognitive biases, which have been extensively documented in the social sciences (Risinger, Saks, Thompson, & Rosenthal 2002). Reflecting on his own experience, John Langbein (1985) has explained that “those of us who serve as expert witnesses are known as ‘saxophones[,]’ . . . a musical instrument on which the lawyer sounds the desired notes.” Courts have criticized experts that altogether “cast aside [their] scholar’s mantle and [became] a shill” for the party that retained them (Mid-State Fertilizer Co. v. Exch. Nat’l Bank of Chi.). The problem is not limited to outliers,
however. In one study, both federal judges and attorneys cited “experts abandon[ing] objectivity and becom[ing] advocates for the side that hired them” as the most frequent problem with expert testimony (Johnson et al., 2000).

Experimental studies have shown that witnesses interviewed by partisan attorneys prior to testifying tend to deliver more biased testimony than those interviewed by nonpartisan attorneys (Sheppard & Vidmar, 1980). In criminal cases, scholars have documented a significant bias of expert opinions in the direction of the side that requested their opinions (Otto, 1989). One litigation-oriented study (Gitlin et al, 2004) compared the opinions rendered by radiologists retained by plaintiffs’ attorneys in asbestos cases to opinions rendered by a panel of “independent” radiologists, retained by defense attorneys. The plaintiffs’ experts found physiological abnormalities 95.9% of the time, while “independent” reviewers found abnormalities only 4.5% of the time. These sorts of biases may operate subconsciously (e.g., an anchoring heuristic driven by the litigant’s suggestion as to the amount of damages) or consciously (e.g., a desire to maintain a stream of income from the litigant).

Expert biases may be the subject of cross-examination in trials, but that is typically a zero-sum game, since both litigants tend to exploit these biases. Although court-appointment of experts is a possible solution, it is almost never actually employed, for cultural and economic reasons, in the adversarial American legal system.

---

33 We put “independent” in scare quotes, because the study authors admit that the research was done on behalf of defense attorneys. See Oliver et al., 2004, arguing that the Gitlin “study was done at the behest of attorneys for defendants in asbestos litigation.”
(Gross, 1991; Cecil & Willging, 1993). Without a real alternative to biased expert
witnesses, it is not practicable for judges to simply exclude all such testimony. Instead,
the exclusion doctrine – based on Daubert v. Merrell Dow Pharmaceuticals – operates
only in marginal cases. It is not a solution for the everyday instances of litigant-induced
biases.

When layperson factfinders (whether judge or jury) are left to rely upon
traditional expert witnesses that are subject to these biases, the factfinders get
relatively weak epistemic signals as to the scientific truth.34 Even where there is a realm
of legitimate disagreement in a field of expertise, the litigation factfinder only sees
handpicked exemplars from each polar extreme of opinion, and are thus unable to
distinguish whether the true distribution of opinion is a 50-50 split or skewed 99-1. This
is a selection bias (Robertson 2010, 184-185).

Jurors “come to understand the adversary system and on the whole evaluate
expert witnesses in the light of this perspective” (Vidmar 1995, 173). When faced with
two counterpoised and seemingly-biased experts, the factfinders sometimes rely on
disparities in credibility to prefer one over the other, but they often feel and sometimes
say that, “the two sides have canceled each other out” (Liptak, 2008). Notwithstanding
the tens of thousands of dollars each side spent on expert witness fees and the hours of
trial time consumed by their testimony, it is as if no expert appeared at all (Brekke and
Borgida, 1988). This epistemic equipoise, resulting from counterpoised hand-picked

34 According to the standard account, the purpose of trials is so that “the truth may be ascertained.” Fed. R.
Evid. 102. See Robertson (2010: discussion surrounding notes 16-26).
experts, may be one reason that jury verdicts result in “discordant outcomes” that contradict the views of independent experts that review the merits of the case (Studdert and Mello, 2007).

As a consequence, to the extent that factfinders are deprived of meaningful and useful expertise, the quality of litigation outcomes is likely very far from the ideal (Studdert and Mello, 2007). The end result is that the deterrence, compensation, and punishment functions of litigation are stymied, and the legitimacy of litigation as a system of dispute resolution is undermined (Solum, 2004, 190).

Over sixty-years ago, biomedical scientists recognized that their objectivity was undermined by extraneous knowledge about which of their subjects were receiving the investigational drug and which were receiving the placebo. Now it goes without saying that, “any process using a human as a perceptor, rater, or interpreter should be ‘as blind as possible for as long as possible’” (Rosenthal, 1978). And in recent years, most scientific journal editors have adopted a blind review process, to ensure that their editors’ decisions are not biased by extraneous information about the author’s identities and institutional affiliations (Snodgrass, 2006). On the other hand, of course, readers of biomedical journal articles are not blinded to the author’s identities, and may rationally use such information in evaluating scientific research, and for allocating their limited time in doing so. For similar reasons, law professors routinely use anonymity in grading (Carrington, 1992).

In recent years, several scholars have been developing the idea that blinding could be used by expert witnesses in criminal and civil litigation. (See Risinger 2009;
Robertson (2010) provides the most extensive exposition of the concept, explaining that in litigation blinding could be implemented through the use of an intermediary between litigant and expert. At the request of a single litigant, an intermediary selects a qualified expert and pays the expert to review the case without knowing which side requested the opinion, thus avoiding selection, compensation, and affiliation biases. The fact that an attorney requested such an opinion would not be disclosed to the court or the adversary.

Under the blind procedure, attorneys would retain ultimate control over their cases, since they individually decide whether to solicit such a blinded expert opinion and then decide whether to call such a blinded expert to trial, after having read his or her expert report. If an opinion turns out to be unfavorable, or not sufficiently strident, an attorney can hide it in “work-product” protection, just as current expert opinions can be hidden. Still, the integrity of the process would be maintained because if an attorney did choose to proceed to trial with a blind expert, he would have to disclose how many blinded experts he consulted on that question, thus allowing the factfinder to evaluate any selection bias. Thus, in practice, litigants would be limited to a single blind expert review on a given question in a case, but would have no risk that an unfavorable blinded opinion could hurt their cases. Robertson (2010, 209-213) has argued that these secrecy and disclosure rules are already part of attorney work-product doctrine, and a blind expert is fully consistent with the rules of evidence, which allows the blind procedure to proceed without any changes to law, without stipulation of party opposite,
4. Effect of blinded experts on juror verdicts

and without intervention of the court. It is a litigant-driven solution, designed to
leverage his or her rational self-interests.

To our knowledge, blinded experts have to date not been put into practice in
litigation. Given that the use of blinded experts seems like it would be a rational
strategy for litigants, why have they not already been employed? Attorneys may be
unsure of whether jurors sufficiently care about the expert’s biases compared to the
facts of the case and the story they can tell about those facts. After all, the case is not
really about the expert witnesses. Attorneys may be unsure of whether jurors would be
able to distinguish between a blinded and an unblinded expert, given that the concept
of blinding is somewhat complicated, and the idea of blinded expert witnesses is novel.
It may be difficult to present the concept in a way that is both understandable and
reassuring to jurors that it is not some sort of sham engineered by the attorneys.

Most fundamentally, attorneys may be unsure whether the jurors would be
more likely to render a judgment in favor of the party employing the blind expert. Even
if attorneys are hopeful about there being some benefit, they may be unsure of whether
the scale of any such benefit would be sufficiently large to offset the cost of buying a
blinded opinion that may turn out to be unfavorable and thus unusable. If the benefit
provided by a blinded expert is relatively small, an attorney may prefer to proceed with
a hand-picked expert, who may also be more skillful as a witness or more malleable in
the details of his testimony.

Further, attorneys may wonder if it would be necessary to first persuade the
judge to provide additional jury instructions explaining and endorsing the concept of
blinded experts for it to have any effect. If an attorney needs that additional boost to get any advantage from the blind procedure, uncertainty about the chances of getting that cooperation may also deter attorneys from using the procedure.

Ultimately, these are empirical questions, ones that turn on whether jurors attend to the problem of litigant-induced biases, whether they can understand the blinding process, and whether they find it a compelling means of removing bias. Some extant research sheds light on these questions. Ivkovic and Hans (2003) have provided an excellent review of the literature on jurors’ perceptions of expert bias. For present purposes, we need only highlight the findings that as jurors evaluate expert testimony they are sensitive to the appearance of partiality. Particularly when jurors have trouble understanding the technical substance of a dispute, they tend to evaluate the credibility of the witnesses as a proxy or heuristic for ascertaining the truth.

In a case study of a complex tort trial, Selvin and Picus (1987, p. 27-28) concluded that, “[c]onfronted with so much complex and confusing information, the jurors tended to evaluate the credibility of these witnesses in large part on their personal characteristics rather than on the information they presented.” On the basis of a case study, Sanders (1993) observed that jurors tended to “discount[ ] all expert opinions as testimony of hired guns,” but noted that jurors still had views about “relative effectiveness of witnesses.” Shuman, Whitaker, and Champagne (1994) performed a survey of jurors, asking them to report what factors they found important in evaluating expert testimony; along with qualifications, familiarity with the case, and quality of reasoning, the jurors said that the appearance of impartiality was an
important factor. Vidmar conducted interviews with jurors after they had deliberated on a case involving expert testimony, and concluded that “when there are competent experts on both sides, and they offer contradictory or confusing opinions, jurors may resolve the differences by relying on general impressions of character and veracity.” (Vidmar, 1995, 172).

In a series of three mock jury experiments using a fairly complex trial stimulus, Cooper and Neuhaus (2000) manipulated the expert witnesses’ pay, credentials, and frequency of testifying. They found (p.156) that the jurors “thought that the plaintiff’s witness was influenced by money as a direct effect of its magnitude,” and the effect was clearest for those experts that testified frequently. Mock jurors found (pp. 165-166, fig.3) for parties with highly paid experts in only nineteen percent of cases and for parties with low paid experts in fifty-seven percent of cases. In a third condition (p. 166), Cooper and Neuhaus found that the interaction becomes especially strong as the complexity of the trial testimony increases. They suggest that in such challenging situations, jurors may shift from central processing to peripheral processing, wherein they begin relying more on cues as to credibility than the substantive testimony itself. Alternatively, Vidmar and Diamond argue (2001) that if jurors had reasons to doubt the messenger’s motives, then the jurors may disregard even that testimony that they understood.

Ivkovic and Hans (2003) studied transcripts and structured interviews with 55 jurors on seven cases to explore how they utilized expert testimony. They “conclude[d] that both the characteristics of the expert (the ‘messenger’) and the substantive and
4. Effect of blinded experts on juror verdicts

stylistic aspects of the testimony itself (the ‘message’) contributes significantly to the overall impact of expert testimony on jurors” (p. 443). In particular, “jurors carefully examined and weighed potential motives for bias” (p. 464). Ivkovic and Hans also conducted a written survey of 269 jurors. Of particular interest here, “seven of ten jurors either agreed or strongly agreed with the statement that ‘lawyers can always find an expert who will back up their client’s point of view, no matter what it is.’ Just one of every ten jurors disagreed with this statement” (p. 452). Men were especially likely to agree with this point (82% versus 64%, odds of 4.62 versus 1.78).

In a 2009 study, Brodsky, Neal, Cramer, and Ziemke asked mock jurors to evaluate the testimony of two actors playing expert witnesses, presenting themselves as either high or low in “likeability,” which they “defined as the degree to which an expert is friendly, respectful, kind, well-mannered, and pleasant.” They found that likeability did in fact impact credibility overall, but that jurors distinguished between that factor and both knowledge and confidence (p. 529). Furthermore, the differences in likeability did not impact jurors’ decisions about whether to sentence the criminal defendant to death versus life without parole.

Brodsky, Griffin, and Cramer (2010) reviewed the research on expert witness credibility and fielded five studies of their own, using mock jurors and videotaped experts to form a “witness credibility scale.” Their factor analysis yielded a scale consisting of 20 paired adjectives loaded onto four factors: “Confidence’ was the strongest factor and explained 49.76% of the variance in expert witness credibility.
Trustworthiness, likeability, and knowledge added significantly to the analysis and accounted for 9.20, 6.56, and 5.10% of the variance, respectively” (p. 899).

Overall then, it is clear that expert testimony can have a significant impact on trial outcomes, and that jurors are sensitive to various factors, including the appearance of partiality. It remains to be seen, however, whether jurors will respond to the concept of blinding expert witnesses, and whether such a response will be significant enough to motivate trial attorneys to utilize that strategy.

Method

Stimulus and design

A two-factor between-subjects design was used, wherein participants watched a video of a medical malpractice trial with two expert witnesses, edited such that neither expert witness was blinded, only the plaintiff’s expert was blinded, or only the defendant’s expert was blinded; moreover, for the two conditions with blinded experts, there either were or were not special instructions from the judge explaining the blind expertise concept. Thus, there were five conditions for equally-weighted randomization. Participants were subsequently asked to render a verdict, make assessments of the credibility of the expert witnesses, and then complete a variety of demographic questionnaires.

The core of the stimulus, viewed by all participants, was a 35-minute video of a staged medical malpractice trial. The script was written by practicing physicians, who also served as both project consultants and the actors playing the expert witnesses. The scenario concerned the failure of a primary care physician to diagnose a possible case of
lumbar radiculopathy and refer the patient to imaging, which allegedly would have allowed timely surgery and avoidance of the permanent disability that the patient now suffers. The primary dispute concerned whether the physician-defendant met the standard of care when, instead of ordering imaging, he simply instructed the patient to take pain-killers and return if the pain got worse. The case was designed so that there was a right answer to this question of medical doctrine, one given by a national practice guideline published in the *Annals of Internal Medicine* (Chou, et al., 2007). According to that guideline and the stipulated facts, the physician *did* violate the standard of care. To avoid confounding with the variables of interest, this guideline was not introduced in the stimulus trial for the present experiment.\(^{35}\) It is only a reference point for analysis.

The trial consisted of the following sequence: the trial judge’s introduction and preliminary instructions (based on the Revised Arizona Jury Instructions (“RAJI”)), very brief opening statements from the plaintiff’s and the defendant’s attorneys, the testimony of plaintiff’s expert, the cross-examination of plaintiff’s expert, the testimony of defendant’s expert, the cross-examination of defendant’s expert, very brief closing statements from the plaintiff’s and defendant’s attorneys, and, lastly, jury instructions from the trial judge (also based on the RAJI). This core video alone constituted the control condition.

\(^{35}\) National practice guidelines are not, of course, used in many other cases involving witnesses, either because they do not exist, are not determinative, or the parties simply decide not to refer to them (Hyams, Lipsitz & Brennan, 1995).
The core video was edited to create four additional conditions. First, approximately ten minutes of extra video footage were distributed across the opening statements, testimony and cross-examinations, a new redirect of one expert, and closing statements, wherein the attorneys and expert witnesses discussed the concept of blind expertise. Using the same language, either the plaintiff’s expert or the defendant’s expert was transformed into a blind expert with this additional material, while the actor and substantive testimony remained the same. For example, the opening statement of the plaintiff’s [defendant’s] attorney, when using the blind expert, contained the following addition:

The evidence will show that our expert, Dr. Pritchard [Dr. Davidson] is a blind expert, like in a blind taste test. This means two things. First, I did not hand-pick Dr. Pritchard [Dr. Davidson] – he comes from a pool of independent experts. Second, I could not have any influence upon him when he made up his mind about this case, since he did not even know which side was asking. That blind protects his integrity and his objectivity. Compare that to the Defendant’s [Plaintiff’s] witness, and then you can decide who to trust.

The blind expert explained the blind procedure during the direct examination from the attorney for his side. The witness explained that he was randomly selected by the American Association for the Advancement of Science, and that he reviewed the case without knowing which side asked. The witness also explained the possible sources of bias (selection, affiliation, and compensation) and noted that blinding was routinely used in biomedical science.

During cross-examination, the attorney using the blind expert interrogated the opposing, non-blind expert witness, insinuating that he was biased, as follows:
Q. So, how many doctors in America do you think could be qualified to review this case?

A. Oh, I don’t know, hundreds, thousands.

Q. Did you ask Mr. Dobbins why he picked you in particular?

A. No.

Q. Do you know why he picked you in particular, rather than hundreds of other doctors?

A. I assume it was on the basis of a recommendation from one of my colleagues, but I don’t know.

Q. Well, it surely wasn’t random, right? Didn’t Mr. Dobbins pick you because he was confident that you’d give him a favorable review of this case?

A. No, I think he just wanted a qualified doctor to review the case, and I reviewed the case and we agreed that it might be reasonable for me to serve.

Note that the non-blind expert firmly denies that he is at all biased, as would be expected in an adversarial trial. Moreover, in the blind expert conditions, the party with the non-blind expert was given the chance to do a re-direct examination, which gave that expert an additional chance to explain that he simply reviewed the facts of the case and rendered an honest opinion:

Q. Doctor, the plaintiff’s attorney has tried to suggest that you’re biased. Just because I asked you [to] come talk to the jurors today. Is that right?

A. That’s right.

Q. But what really informed your decision on this case?

A. I read the report. I read the medical record. I read the report, the complaint from the plaintiff. I talked with the defendant. And, you know, based on my own experience and expertise, I put together my report.

Q. And, I didn’t tell you what to put in your report, did I?

A. No, you didn’t.
Q. I didn’t tell you what to say today, did I?
A. No, you didn’t.

Q. Doctor, can you explain to the jury about professionalism and what that means?
A. Sure, I mean, you know, physicians have to adhere to certain standards that are required of members of a professional discipline who have, you know, credentials certified by the state and by the, you know, boards of professional conduct, and, you know, so you have to act with a certain amount of reasonableness and truthfulness.

Q. And have you done that here today?
A. Yes, to the best of my ability.

Q. Now, are there certain codes or rules that you have to follow as a professional?
A. Yes, you mean, there are certain ethical codes and basic standards of conduct.

Q. And are you complying with those in your testimony today?
A. Yes.

Q. Thank you.

In all versions of the video, the judge provided several minutes of general jury instructions, defining medical negligence and other relevant concepts, including a standard charge for the jury to consider the credibility of witnesses. We worried that the attorneys’ questions and argument about blinding may not be sufficient to change jurors’ votes significantly. Thus, for two of the experimental conditions, the videos were further edited to add a short segment of special jury instructions from the judge regarding the blind expert, to test the hypothesis that this further intervention may make blind expertise more efficacious. In particular, the following language was added:
I previously instructed that you should consider the potential bias of the witnesses you heard. You have heard testimony from a certain kind of expert witness called a “blind expert,” which is a method used to try to minimize bias. In this case, the Plaintiff’s [Defendant’s] expert, Dr. Pritchard [Dr. Davidson] was blinded. This expert was randomly selected by a neutral third-party from a pool of qualified experts, and [he] rendered an initial opinion about this case without being influenced by either side. Therefore you may find the blind expert more credible and you may decide to give the testimony of that expert additional weight. However, the ultimate decision about the credibility of witnesses is yours.

The end result of all this editing was five video conditions: a control containing the core video stimulus; two videos wherein the plaintiff’s expert was a blind expert, either with or without special judge’s instructions about blinded experts; and two videos wherein the defendant’s expert was a blind expert, either with or without special judge’s instructions about blinded experts. The full videos and transcripts of each condition are available upon request from the authors.

Instrument

As noted above, we collected demographic data before showing the trial video to the participants. After the videos, binary verdict judgments were elicited by asking: “Based on the instructions provided by the judge in the video, do you believe that the Plaintiff has proved, by the greater weight of the evidence, that the Defendant committed medical negligence?” Participants responded “yes” or “no.”

We also asked respondents: “Based on the evidence you saw, please rate this case on a scale of 1 to 6,” where 1 was defined as “clearly not medical negligence” and 6 was defined as “clearly medical negligence.” This scalar variable was used for more sensitive data analyses and as a proxy for verdict certainty, in that more certain jurors
presumably respond toward the extremes, while equivocal jurors respond towards the middle.

Participants who found medical negligence were further asked how much money the plaintiff should be compensated for pain and suffering. Participants were not asked to decide economic damages, because the trial did not present evidence about such damages (for the sake of conserving time, and on the assumption that damages may be stipulated by the parties in such cases). Participants were also asked to type a sentence or two explaining their decisions.

**Credibility assessments**

To assess the relative credibility of each witness, subjects were asked whether they agreed that the expert witness was knowledgeable, logical, clear, honest, trustworthy, and fair. For each adjective, there was a four-point labeled Likert rating (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree). For analyses, these factors were grouped into two composites; how and why this was done is described more fully below.

**Participants and randomization**

Jury-eligible adults were recruited for an online mock jury experiment via Amazon Mechanical Turk, for three dollars compensation, beginning September 14, 2010. (This platform is increasingly used for social science research, and has been validated by comparison to known results; see Paolacci, Chandler & Ipeirotis, 2010.) Participants were told that the study concerned a mock trial, but were not informed
about the hypotheses being tested and were blinded as to their assignment into experimental conditions.

Three hundred and sixty-four persons consented to participate within the first 48 hours. Fifty-eight persons exited the study before completion, and another 31 persons were excluded for failure to follow instructions (e.g., they completed the study in under 35 minutes, indicating that the 35+ minute trial video was not watched in its entirety). As shown in Table 1, the remaining 275 participants, who constitute the overall sample for analyses purposes, included 195 females and 80 males, aged between 18 – 77 years ($M = 32.5$ years, $SD = 10.9$ years). The most common ethnicities were Caucasian (78.2%), African American (8.7%), and Asian (4.4%). As for formal education, 2.2% reported they had not completed high school, 17.8% had a high school degree, 37.5% reported some college, 28.4% were college graduates, and 14.1% had post-graduate work. All persons consented to participate according to Institutional Review Board standards.

Each participant was then randomly assigned to one of the five experimental video conditions. As shown in Table 1, randomization was successful, with no statistically significant differences on demographic variables across experimental conditions.
Table 1. Demographics of Experimental Subjects. The sample was somewhat more highly educated, more often female, and younger than the U.S. population, but had a similar racial distribution (white/non-white) and the differences were successfully randomized across the five experimental conditions.

<table>
<thead>
<tr>
<th>Education</th>
<th>Neither BE (n = 61)</th>
<th>BE for Plaintiff (n = 114)</th>
<th>BE for Defendant (n = 100)</th>
<th>Subject Totals (N = 275)</th>
<th>U.S. Census</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Instructions</td>
<td>Instructions</td>
<td>Subtotal</td>
<td>No Instructions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(n = 54)</td>
<td>(n = 60)</td>
<td>(n = 52)</td>
<td>(n = 48)</td>
<td></td>
</tr>
<tr>
<td>&lt; HS Diploma/GED</td>
<td>2%</td>
<td>3%</td>
<td>3%</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>HS Diploma/GED</td>
<td>16%</td>
<td>19%</td>
<td>17%</td>
<td>17%</td>
<td>18%</td>
</tr>
<tr>
<td>Some College/Assoc.</td>
<td>43%</td>
<td>37%</td>
<td>38%</td>
<td>35%</td>
<td>38%</td>
</tr>
<tr>
<td>College Grad</td>
<td>27%</td>
<td>32%</td>
<td>38%</td>
<td>35%</td>
<td>32%</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>13%</td>
<td>9%</td>
<td>6%</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>30%</td>
<td>25.4%</td>
<td>33%</td>
<td>29%</td>
<td>49%</td>
</tr>
<tr>
<td>Female</td>
<td>71%</td>
<td>74.6%</td>
<td>67%</td>
<td>71%</td>
<td>51%</td>
</tr>
<tr>
<td>Age Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>28%</td>
<td>21.1%</td>
<td>31%</td>
<td>26%</td>
<td>24%</td>
</tr>
<tr>
<td>25-34</td>
<td>51%</td>
<td>42.1%</td>
<td>35%</td>
<td>36%</td>
<td>41%</td>
</tr>
<tr>
<td>35-44</td>
<td>12%</td>
<td>17.5%</td>
<td>15%</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>45-59</td>
<td>10%</td>
<td>19.3%</td>
<td>15%</td>
<td>17%</td>
<td>16%</td>
</tr>
<tr>
<td>60+</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>73%</td>
<td>75%</td>
<td>85%</td>
<td>84%</td>
<td>78%</td>
</tr>
<tr>
<td>Non-White</td>
<td>26%</td>
<td>25%</td>
<td>15%</td>
<td>16%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Note: “BE” is an abbreviation for blind expert, and “instructions” refers to a special jury charge on the blind expert concept.
Results

Overall regression on verdict

We used binary logistic regression to assess the impact of the experimental manipulations on the likelihood that a participant would render a verdict of medical negligence, while also controlling for demographic variables of gender, age, race, and education. The model contained the following independent variables: Plaintiff BE (whether or not the plaintiff had a blind expert), Defendant BE (whether or not the defendant had a blind expert), Instructions (whether or not the judge provided special jury instructions focusing on blinding, in addition to the standard instructions), subject’s gender, age, race (non-white or white), and education (less than college degree or at least college degree). A hierarchical method of entry was used. The four demographic variables were force entered in the first step, followed by forced entry of Plaintiff BE and Defendant BE in the second step, and then forced entry of Instructions in the final step.37

The full model containing all predictors was statistically significant, \( \chi^2 (8) = 26.83, p = .001 \), indicating that the model was able to successfully distinguish between

36 The use of instructions, if it increases the credibility of the blind expert as predicted, would increase or decrease the likelihood of a negligence verdict, depending on whether the plaintiff or defendant, respectively, were the side with the blind expert. As such, the main effect of Instructions would potentially cancel itself out, thereby revealing a misleading null result. To avoid this possibility, the model included the interaction terms of Instructions*Plaintiff BE and Instructions*Defendant BE.

A different approach is to include the Instructions term but restrict the logistic regression sample to compare the condition with no blind expert to either Plaintiff BE or Defendant BE, but not both simultaneously – thus meaning that the Instructions effect relates to Plaintiff BE or Defendant BE, respectively. The results are consistent with those reported in Table 2 with the interaction term.

37 Thus, using this method of entry, the demographic variables are used first, to explain as much variation in outcomes as possible; after that, the experimental variables are used to account for any remaining variance.
The model correctly classified 63.3% of cases and, with effect size estimates in the range of .09 (Cox and Snell $R^2$) to .12 (Nagelkerke $R^2$), constitutes a meaningful result. Only Plaintiff BE and Defendant BE made a unique, statistically significant contribution to the model. (See Table 2.) In particular, the odds of a verdict in favor of the plaintiff more than doubled when the plaintiff’s expert was a blind expert (odds ratio = 2.18, $p = .048$). Likewise, the odds of a verdict in favor of the defendant more than doubled when the defendant’s expert was a blind expert (odds ratio = 0.46, $p = .059$). These findings affirm our primary hypothesis that the presence of a blind expert would have a large effect on the outcome of trials.

---

38 Regression results are materially the same when using robust standard errors and classical standard errors (the latter are reported herein). See King & Roberts (2012) (recommending models be tested to ensure that both types of error terms generate the same results, rather than simply reporting robust standard errors).
Table 2. Logistic Regression Predicting the Odds of a Plaintiff Verdict. The overall model was statistically significant and the presence of a blind expert more than doubled the odds of winning, for either side. Neither the presence of special jury instructions nor any covariates made a unique, statistically significant contribution to the model.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>S.E.</th>
<th>Wald’s $\chi^2$</th>
<th>$p$</th>
<th>Lower Odds Ratio</th>
<th>Upper Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.003</td>
<td>0.012</td>
<td>0.047</td>
<td>0.828</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Gender</td>
<td>0.305</td>
<td>0.286</td>
<td>1.142</td>
<td>0.285</td>
<td>0.78</td>
<td>1.36</td>
</tr>
<tr>
<td>Race</td>
<td>-0.423</td>
<td>0.313</td>
<td>1.831</td>
<td>0.176</td>
<td>0.36</td>
<td>0.66</td>
</tr>
<tr>
<td>Education</td>
<td>-0.309</td>
<td>0.260</td>
<td>1.406</td>
<td>0.236</td>
<td>0.44</td>
<td>0.73</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>0.778</td>
<td>0.393</td>
<td>3.921</td>
<td>0.048*</td>
<td>1.01</td>
<td>2.18</td>
</tr>
<tr>
<td>Defendant BE</td>
<td>-0.776</td>
<td>0.410</td>
<td>3.579</td>
<td>0.059*</td>
<td>0.21</td>
<td>0.46</td>
</tr>
<tr>
<td>Instructions*Plaintiff BE</td>
<td>-0.167</td>
<td>0.395</td>
<td>0.180</td>
<td>0.672</td>
<td>0.39</td>
<td>0.85</td>
</tr>
<tr>
<td>Instructions*Defendant BE</td>
<td>0.430</td>
<td>0.438</td>
<td>0.960</td>
<td>0.327</td>
<td>0.65</td>
<td>1.54</td>
</tr>
<tr>
<td>Constant</td>
<td>0.130</td>
<td>0.498</td>
<td>0.068</td>
<td>0.794</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* $p < .05$ or marginally significant.

Note: $R^2 = .09$ (Cox & Snell), .12 (Nagelkerke). Model $\chi^2 (8) = 26.83, p = .001$.

On the other hand, the regression model also shows that the use of special jury instructions did not have a significant effect on verdict, and this was true whether the plaintiff ($p = .67$) or defendant ($p = .33$) retained the blind expert. This finding disconfirms our secondary hypothesis that this additional intervention by the judge would magnify the effect of the blind expert. Instead it appears that a purely litigant-driven blinding intervention will suffice.

Because randomization succeeded in distributing covariates across conditions, and these overall regression results show that covariates are not driving our main results, we proceeded to conduct planned comparisons of individual experimental conditions, reported in the following sections. Because comparisons of central
tendencies (means and medians) may be easier to interpret and exhibit than regression models, we present such comparisons of primary dependent variables including verdict and verdict certainty, damages, and expert credibility, using parametric and non-parametric tests of significance.

**Verdict and certainty**

When neither side had a blind expert – as in a usual trial – most participants rendered a verdict in favor of the physician-defendant, even though the scenario had been designed to be a case of real medical malpractice that the plaintiff should have won. In this control condition, 46% of the jurors found medical negligence. (See Table 3). Introduction of a blind expert significantly altered this status quo. When a blind expert appeared for the plaintiff (and without special jury instructions), 65% of the jurors found negligence, an increase of 19% compared to the Neither BE control condition. A Pearson’s chi-square test reveals this difference to be significant \( \chi^2 (1) = 4.14, p = .042 \). When a blind expert appeared for the defendant (and without special jury instructions), only 27% of the jurors found negligence, a decrease of 19% compared to the Neither BE control condition. This difference is also significant \( \chi^2 (1) = 3.62, p = .037 \). The planned comparison thus further confirmed our primary hypothesis that blind experts would drive outcomes favorable to their sponsors.
**Table 3.** Frequency (and Percentage) of Verdicts in Favor of the Defendant and Plaintiff, along with Mean Certainty Score on 6-Point Scale (and Standard Deviation) by Blind Expert (BE) and Instruction Conditions. Compared to the control condition where 46% of respondents voted for the plaintiff, the presence of a blind expert for the plaintiff increased that rate to 62% (a difference of 16 points, \( p = .042 \)) and a blind expert for the defendant decreased that rate to 31% (a difference of 15 points, \( p = .037 \)). Special jury instructions from the judge were associated with statistically insignificant decreases in win rates.

<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>Verdict for Plaintiff (%)</th>
<th>Mean Certainty (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE (control)</td>
<td>61</td>
<td>28 (45.9)</td>
<td>3.46 (1.51)</td>
</tr>
<tr>
<td>BE for Plaintiff</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>54</td>
<td>35 (64.8)</td>
<td>4.11 (1.44)</td>
</tr>
<tr>
<td>Instructions</td>
<td>60</td>
<td>36 (60.0)</td>
<td>3.90 (1.57)</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>71 (62.3)</td>
<td>4.00 (1.51)</td>
</tr>
<tr>
<td>BE for Defendant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>52</td>
<td>14 (26.9)</td>
<td>2.90 (1.49)</td>
</tr>
<tr>
<td>Instructions</td>
<td>48</td>
<td>17 (35.4)</td>
<td>3.04 (1.71)</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>31 (31.0)</td>
<td>2.97 (1.60)</td>
</tr>
</tbody>
</table>

As in the logistic regression, our secondary hypothesis about jury instructions was not confirmed in this planned comparison across conditions. Within the blind expert conditions, the use of jury instructions about the blind expert was not associated with increased win rates for the party bringing the blind expert. In fact, the jury instructions were associated with statistically insignificant decreases in win rates (a 5% difference when the plaintiff had the blind expert (\( \chi^2 (1) = 0.28, p = .596 \)), and an 8% difference when the defendant had the blind expert (\( \chi^2 (1) = 0.84, p = .359 \))).

The six-point scale for how certainly the jurors found medical negligence reveals that, in the presence of a blind expert, jurors had stronger opinions about how “clearly” the defendant committed medical negligence. The mean score in the control condition was 3.46. (See Table 3). This score was higher at 4.00 (indicating more clearly
malpractice) when the plaintiff retained a blind expert, and was lower at 2.97 (indicating more clearly not malpractice) when the defendant retained a blind expert. These results were statistically significant, given an ordinal regression, \( \chi^2 (4) = 23.6, p < 0.001. \)

**Monetary damages and hedging thereof**

Participants who found medical negligence also determined how much money the plaintiff should be awarded in pain and suffering damages. Recall that participants were not asked to decide economic damages (e.g., medical bills and lost wages). Also, the attorneys did not suggest specific values for pain and suffering awards. Descriptive statistics for damages are listed in Table 4. The data were severely right-skewed, with several far outliers. These outliers might be explained by the fact that jurors were not given any guidance or limitations as to the amount of pain and suffering damages they could award.

---

39 An ordinal regression was performed rather than a t-test, because the latter presumes interval scores, which are not present in a Likert-scale.

40 In the control condition, the \( z \)-score for skewness equaled 51.53, which is highly significant \( (p < .0001) \); the skewness \( z \)-scores were also highly significant in the BE for Plaintiff \( (z_{skew} = 30.63, p < .0001) \) and BE for Defendant \( (z_{skew} = 23.66, p < .0001) \) conditions.

41 In the real world of litigation, there is heterogeneity as to whether states allow attorneys to request specific amounts of damages for pain and suffering, and furthermore there is heterogeneity as to whether attorneys actually do so when allowed. See Kahneman Schkade, & Sunstein, (1998) for a general discussion of the difficulty jurors face in mapping judgments of pain and suffering into a numeric dollar amount. See Diamond, Rose, Murphy, & Meixner (2011), for a discussion of whether attorneys are allowed to give reference points for pain and suffering damages, how often attorneys do so when allowed, and how juries react thereto when they are offered. See especially, pp. 6-7, explaining that a few states, such as New Jersey and Pennsylvania, forbid attorneys from asking for a specific amount for pain and suffering, because such requests would be of “an arbitrary amount” and thus “highly improper,” and p. 20, showing that in a sample of Arizona cases, where attorneys are allowed to name a dollar figure for pain and suffering, they did so in 21 out of 31 cases (68%), and defense attorneys conceded some amount of economic damages (contingent on a finding of liability) in 19 out of 33 cases (58%). Thus, our trial stimulus – which stipulated economic damages, and thus did not ask the jury for a finding on that item, but asked for a pain and suffering award, without providing a reference point, is not an altogether uncommon scenario. Still, for future experiments, it may be prudent to provide such a reference point, as a way to reduce variance in the data.
Table 4. Mean and Median Pain and Suffering Awards including Zeros for Defense Verdict (U.S. Dollars), by Blind Expert (BE) and Instruction, with Standard Deviation and 95% Confidence Interval. The mean award was higher (p = .02) or lower (p = .01) than in the control condition by about $163,000, depending on whether the plaintiff or defendant, respectively, retained the blind expert

<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE</td>
<td>61</td>
<td>278,313</td>
<td>10,000</td>
<td>780,584</td>
<td>±199,917</td>
</tr>
<tr>
<td>BE for Plaintiff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>54</td>
<td>370,018</td>
<td>100,000</td>
<td>613,708</td>
<td>±167,510</td>
</tr>
<tr>
<td>Instructions</td>
<td>60</td>
<td>505,695</td>
<td>72,500</td>
<td>1,338,990</td>
<td>±345,897</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>441,427</td>
<td>100,000</td>
<td>1,057,069</td>
<td>±196,144</td>
</tr>
<tr>
<td>BE for Defendant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>52</td>
<td>86,599</td>
<td>0</td>
<td>216,363</td>
<td>±60,235</td>
</tr>
<tr>
<td>Instructions</td>
<td>48</td>
<td>144,583</td>
<td>0</td>
<td>481,914</td>
<td>±139,933</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>114,431</td>
<td>0</td>
<td>367,722</td>
<td>±72,964</td>
</tr>
</tbody>
</table>

The presence of a blind expert impacted pain and suffering damages, but the presence of special jury instructions on blinding did not further increase that effect.

Compared to the $278,313 awarded in the control condition on average, the blind expert for the plaintiff yielded an additional $163,000, totaling $441,427.

Symmetrically, when the defendant brought the blind expert, the jury awarded $163,000 less on average, totaling $114,431. These effects are notable, especially since the expert witnesses never directly discuss pain and suffering damages in any of the experimental conditions. Because the data were so significantly skewed, a non-parametric Mann-Whitney test was used. Plaintiffs received significantly more pain and suffering damages when they retained a blind expert (Median = 100,000) relative to the control (Median = 10,000), \( U = 2,755, z = -2.31, p = .02, r = -.17 \). Likewise, defendants paid
significantly less pain and suffering expenses (\(Mdn = 0\)) when they retained a blind
expert relative to the control (\(Mdn = 10,000\)), \(U = 2,379, z = -2.55, p = .01, r = -.20\).

These strong effects, of course, are being driven in large part by the underlying
win-rates, in particular the zero damages award that comes with a finding that the
defendant is not-negligent. Still, the finding highlights the dramatic effect that using a
blind expert can have on the economic bottom line for a litigant.

A separate question, however, is whether the blinded expert impacted damages
awards, aside from the differential win rates. We did not hypothesize such an effect,
since the expert testimony in this case shed light on the question of whether the
physician met the standard of care (and thus the likelihood of defense verdict), not the
severity of the plaintiff’s injury. Nonetheless, to investigate the possibility that the type
of expert impacted pain and suffering awards, Table 5 shows the damages awarded in
each condition, using data only from those who rendered a verdict in favor of the
plaintiff. Even with the defense verdicts excluded, the mean changes appear
substantial: plaintiffs with a blind expert received about $90,000 more, and defendants
with a blind expert essentially “lost better” – paying about $235,000 less. However,
Mann-Whitney tests revealed that the differences were not statistically significant at the
.05 level.\(^{42}\) Especially given the high variance in this data, we were not sufficiently
powered to test the hypothesis that the presence of a blind expert will impact the

\[^{42}\] Neither BE against BE for Plaintiff: \(U = 873, z = -.94, p = .346\). Neither BE against BE for Defendant: \(U = 365, z = -1.05, p = .293\). Data in all three conditions were again highly skewed: Control \(z_{\text{skew}} = 7.17, p < .0001\); BE for Plaintiff \(z_{\text{skew}} = 20.32, p < .0001\); BE for Defendant \(z_{\text{skew}} = 7.67, p < .0001\).
amount of non-economic damages awarded.\textsuperscript{43} Future experimenters could, however, explore this dynamic with a larger sample size.

**Table 5.** *Pain and Suffering Awards of those who Found Negligence (U.S. Dollars).* Damages awards varied widely, with apparent changes on mean and median in expected directions (higher when the plaintiff had a blind expert and lower when the defendant did), but not at statistically significant levels.

<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE</td>
<td>28</td>
<td>588,214</td>
<td>225,000</td>
<td>1,782,276</td>
</tr>
<tr>
<td>BE for Plaintiff</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>35</td>
<td>551,600</td>
<td>300,000</td>
<td>696,269</td>
</tr>
<tr>
<td>Instructions</td>
<td>36</td>
<td>800,847</td>
<td>500,000</td>
<td>1,667,209</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>677,978</td>
<td>300,000</td>
<td>1,281,020</td>
</tr>
<tr>
<td>BE for Defendant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Instructions</td>
<td>14</td>
<td>286,946</td>
<td>125,000</td>
<td>346,574</td>
</tr>
<tr>
<td>Instructions</td>
<td>17</td>
<td>408,235</td>
<td>100,000</td>
<td>753,558</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>353,459</td>
<td>100,000</td>
<td>598,888</td>
</tr>
</tbody>
</table>

The variability in the awards was also seemingly affected; in particular, the presence of a blind expert reduced the amount of variability, with the standard deviation being about a half million dollars less than when the plaintiff’s expert was blinded, and just over a million dollars less when the defendant’s expert was blinded.

These observations were, however, short of statistical significance.\textsuperscript{44} Nonetheless, such apparent changes in variance are noteworthy. Scholars have observed a similar pattern

---

\textsuperscript{43} We confirmed the power calculations provided by an anonymous reviewer, showing that the estimated power is 0.06 and 0.10 for the two conditions comparisons with the control group, which means that we had less than 10\% chance of obtaining a statistically significant effect of the observed magnitude, given the sample sizes.

\textsuperscript{44} A statistical test of variability can be computed by using deviation scores, wherein one calculates the absolute distance between each data point and its mean, and then runs a $t$-test on those deviation scores. See Saks et al. (1997), p. 250. Using this method to test Neither BE against BE for Plaintiff and then against BE for Defendant reveals, respectively, $t(97) = 0.05$, $p = .96$ and $t(57) = 1.51$, $p = .14$. 

of “horizontal inequity” across real cases in which plaintiffs with similar injuries received vastly different awards, and have suggested reforms to reduce such variability (e.g., Saks et al., 1997, and Kahneman, Schkade & Sunstein, 1998). With a very high standard deviation observed in the control condition (nearly triple the mean), the present experiment replicates this problem in a setting where the plaintiff’s injury and circumstances are identical. If confirmed in a larger study, this finding would be contrary to the suggestions that unobserved variation in the plaintiff’s circumstances may explain jury variance (Hans & Eisenberg, 2011, p. 380).

A more realistic assessment of the impact of such juror decisions on the economic value of cases requires several adjustments to the foregoing data. First, it is likely that during the jury deliberation process, and then again in review of the verdict by the trial and appeals judges using the remittur power and statutory caps on non-economic damages, extreme individual assessments of liability would yield to more modest judgments. To approximate this moderating dynamic, extreme outliers were transformed to be within two standard deviations from the mean. (See Saks et al, 1997, p. 249, justifying the use of a two-standard deviations transformation of damages awards over logarithmic transformation or truncation). Second, we assumed that, on average, a medical malpractice case tried to a jury would be worth about $500,000 in

---

45 See Davis (1996) and Diamond & Casper (1992), showing that the best predictor of a jury damages award is the median of individual awards, which necessarily reduces the effect of outliers, and Snyder (2000), p.320 explaining that use of remittitur has become “pervasive in American law ... as an all-purpose effort to ... reduce exorbitant damage awards, and AMA 2011, p. 1. showing that “close to 30 states have laws in place that limit damages in medical liability actions.”
economic damages (primarily for medical expenses and lost earnings), in addition to whatever was awarded in pain and suffering damages. This is a conservative assumption.\footnote{Given the fixed costs associated with prosecuting a case, plaintiff’s attorneys will often decline cases that have modest potential damages. A review of 257 California medical malpractice cases showed mean economic damages awarded by juries of $950,000 and mean non-economic damages awards of $687,000, for a total of $1,637,000 on the mean. (Pace et al., 2004, p. 20, Table 3.1). After imposition of statutory damages caps, the final mean judgments were $950,000 for economic damages (nearly double what we assumed here), $200,000 for non-economic damages (about the same as what we found here after adjusting the outliers), equaling $1,150,000. (Ibid. at 23, Table 3.3) Comparing $1,150,000 to the $709,984 mean simulated award in the control condition here, as shown in Figure 1, shows that altogether our assumptions were conservative. If we had assumed a larger amount of economic damages, or been less aggressive in policing outliers in the non-economic damages, we could have demonstrated an even larger economic value for the blind procedure.} Finally, we included defense verdicts as zero damages, because these would be of primary interest to litigants on both sides. Altogether then, the resulting figure represents the economic value of the case—the overall loss suffered by the defendant going to trial under these conditions or the overall benefit enjoyed by the plaintiff. Aside from litigation costs (not calculated here), this number would represent a potential settlement value for the case.
Figure 1. Simulation of Economic Value of Case (US Dollars) When Neither Side (No BE), Only the Plaintiff (BE Pl.), or Only the Defendant (BE Def.) Has a Blind Expert, Including Defense Verdicts as Zeros. Outlier award values were transformed to within 2 standard deviations, and $500,000 economic damages were assumed. On these assumptions, the tactic of using a blind expert pays over $100,000 on average to the litigant that uses the tactic, conditional on the expert rendering a favorable, usable opinion not rebutted by a blind expert on the other side.

<table>
<thead>
<tr>
<th></th>
<th>BE Pl.</th>
<th>No BE</th>
<th>BE Def.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$854,408</td>
<td>$709,984</td>
<td>$603,591</td>
</tr>
<tr>
<td>median</td>
<td>$572,500</td>
<td>$0</td>
<td>$0</td>
</tr>
</tbody>
</table>

Figure 1 depicts these results, collapsing across the instructions and no-instructions conditions. The shift in over $100,000 in case value for whichever side brings a blind expert shows that blind experts may be a rational strategy for litigants, and thus a practical reform to the system of civil litigation. When the plaintiff had a blind expert, total awards were significantly greater, relative to the control condition, by about $145,000 on the mean, and $572,000 on the median, $U = 2,773, z = -2.30, p = .02$. Likewise, awards were significantly less than in the control condition, by about $106,000 on the mean (with identical medians of $0), when the defendant had a blind expert, $U =$
2,533, \( z = -2.09, p = .037 \). These results are driven by both the significant differences in win rates across conditions (see Table 4), as well as the large but statistically insignificant differences in the amounts awarded across conditions (see Table 5).

The difference in mean values will be important to rational litigators who are deciding whether to invest in a blind expert opinion ex ante, but the differences in medians may also be particularly attractive given the aversion to losing. In the control condition, representing status quo litigation, most cases are economic losers that are only subsidized by the rare case that wins big. With a blind expert, the median case has positive economic value.

Figure 2 shows mean and median pain and suffering awards, at each level of the jurors’ assessments of how “clearly” medical negligence was proven, in all experimental conditions. Note that only those finding negligence are included (thereby censoring levels 1, 2, and 3, where no damages would be awarded, regardless of clearness), and that outliers were transformed to within two standard deviations of the group mean. As can be seen, as verdict certainty increased, the mean and median monetary award also increased. A Kruskal-Wallis test (used because the data was highly skewed) revealed these differences to be significant, \( \chi^2 (4, N = 130) = 10.694, p = .030 \).
In principle, in cases where the malpractice was in fact more egregious, a plaintiff could experience more pain and suffering, which would thus appropriately support a more generous damages award. Our experiment did not, however, manipulate any substantive testimony about how egregious the malpractice may have been; we simply manipulated the credibility of the witnesses, through use of blinded experts. Thus, we think that a more plausible explanation is that jurors responded on the verdict certainty scale in terms of their own epistemic uncertainty about whether malpractice had occurred, one way or the other. They then apparently “hedged” any uncertainty in their verdict by awarding a smaller pain and suffering award than if they
were confident it was clearly medical malpractice, where they presumably awarded what they perceived as full compensation. Given that the presence of a blind expert also affected juror certainty levels, this hedging dynamic may explain why the presence of blind expert was correlated with damages awards (see Figure 1).

**Expert credibility assessments**

We hypothesized that use of a blind expert would impact verdict outcomes because jurors would perceive those experts as more credible. In this section, we examine overall credibility across conditions, and create two subsidiary constructs to investigate the dynamic more precisely.

Research on perceptions of credibility suggests that credibility is usefully thought of as a construct composed of multiple factors. For example, Brodsky, Griffin & Cramer (2010) propose a Witness Credibility Scale composed of the factors *knowledge*, *likeability*, *trustworthiness*, and *confidence*. In this study, participants judged each witness in terms of how *knowledgeable, logical, clear, honest, trustworthy, and fair* he seemed, using a 4-point Likert scale from 1 (strongly disagree) to 4 (strongly agree) for each term. We added the scores on each to create an overall credibility score.

As shown in Table 5, assessments of overall credibility followed the expected pattern. Under the usual circumstances when neither expert was a blind expert (that is, the Neither BE condition), even though the defendant tended to win, the plaintiff’s expert was perceived as more credible ($M = 6.39, SD = 1.21$) than the defendant’s expert ($M = 5.96, SD = 1.28$), $t(61) = -2.03, p = .05, r = .25$ (see Table 6). This discrepancy in credibility was significantly greater whenever the plaintiff’s expert was a blind expert
4. Effect of blinded experts on juror verdicts

\(M_{\text{plaintiff}} = 6.94, SD = 1.17, \text{ versus } M_{\text{defendant}} = 5.35, SD = 1.20\), \(t(114) = -9.48, p < .001, r = .66\). On the other hand, when the defendant’s expert was a blind expert, the defendant’s expert was perceived as significantly more credible \(M = 6.26, SD = 1.07\) than the plaintiff’s expert \(M = 5.73, SD = 1.14\), \(t(89) = 3.43, p = .001, r = .34\).

Table 6. Mean (Standard Deviation) assessments of overall expert credibility, on 4-point Likert scales from 1 (strongly disagree) to 4 (strongly agree) for each term, subdivided by skill and genuineness, when neither (Neither BE), only the plaintiff (BE for Plaintiff), or only the defendant (BE for Defendant) retains a blind expert. Transforming one expert into a blind expert not only increased his own credibility, but also decreased the other expert’s credibility.

<table>
<thead>
<tr>
<th>Blind Expert</th>
<th>Plaintiff’s Expert</th>
<th>Defendant’s Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Credibility</td>
<td>Skill</td>
</tr>
<tr>
<td>Neither BE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.21)</td>
<td>(.62)</td>
</tr>
<tr>
<td>BE for Plaintiff</td>
<td>6.94*</td>
<td>3.51*</td>
</tr>
<tr>
<td></td>
<td>(1.17)</td>
<td>(.58)</td>
</tr>
<tr>
<td>BE for Defendant</td>
<td>5.73*</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>(1.14)</td>
<td>(.53)</td>
</tr>
</tbody>
</table>

* Planned contrast relative to Neither BE condition is significant at \(p < .05\).

NOTE: Based on four-point Likert scales from 1 (strongly disagree) to 4 (strongly agree) for each term, subdivided by skill and genuineness, when neither (Neither BE), or the plaintiff (BE for plaintiff), or only the defendant (BE for defendant) retains a blind expert. Transforming one expert into a blind expert not only increased his own credibility, but also decreased the other expert’s credibility.

Interestingly, transforming one expert into a blind expert not only increased his own credibility, but it also decreased the other expert’s credibility. For example, the credibility of the defendant’s expert dropped significantly from a 5.96 (in the Neither BE condition) to a 5.35 (in the BE for Plaintiff condition), when the plaintiff had a blind
Likewise, the plaintiff’s expert’s credibility dropped from a 6.39 (Neither BE) to a 5.73 (BE for Defendant).

Since blinding an expert both enhanced his credibility and harmed the credibility of the opposing expert, the end-result was an even larger gap in the credibility between the two witnesses. There was only a 0.43 difference in the credibility of the two witnesses in the Neither BE condition, but this increased to 1.59 and 0.55 credibility gaps, respectively, in the BE for Plaintiff and BE for Defendant conditions.

To investigate the dynamic more closely, the overall credibility score can be disaggregated into its ‘skill’ and ‘genuineness’ components, which parallel the knowledge and trustworthiness factors posited by Brodsky, Griffin, & Cramer (2010). The ‘skill’ construct, consisting of knowledgeable, logical, and clear, was meant to capture the expertise of the witness in terms of, for example, his technical training or intellect. The ‘genuineness’ construct, in contrast, consisted of honest, trustworthy, and fair, and was meant to capture the believability of the witness. Skill and genuineness

---

47 Whether the defendant’s expert witness was a blind expert had a significant effect on his credibility as perceived by participants, $F(2, 272) = 17.33, p < .001$. Planned contrasts revealed a significant increase in overall credibility in the Defendant BE condition ($M = 6.28, SD = 1.07$) relative to the Neither BE condition ($M = 5.96, SD = 1.28$), $t(272) = -1.69, p = .045$ (1-tailed), $r = .10$. When only the plaintiff used a blind expert, the defendant’s expert witness suffered a worse credibility assessment ($M = 5.35, SD = 1.20$) than in the Neither BE condition, $t(272) = -3.27, p < .001$ (1-tailed), $r = .19$.

48 Whether the plaintiff’s expert witness was a blind expert had a significant effect on his credibility as perceived by participants, $F(2, 272) = 28.64, p < .001$. Planned contrasts revealed a significant increase in overall credibility of the plaintiff’s expert in the Plaintiff BE condition ($M = 6.94, SD = 1.28$) relative to the Neither BE condition ($M = 6.94, SD = 1.17$), $t(272) = 2.96, p < .001$ (1-tailed), $r = .18$. When only the defendant used a blind expert (Defendant BE), the plaintiff’s expert witness suffered a worse credibility assessment ($M = 5.73, SD = 1.14$) than in the Neither BE condition, $t(272) = 3.50, p < .001$ (1-tailed), $r = .21$.

49 To appreciate the possible difference between these two factors, you might imagine, for example, an expert with outstanding credentials (but who seems to be lying). This expert would have a high skill score (he or she could give an accurate opinion), but a low genuineness score (since he lies). On the other hand, an expert might appear to be doing his honest best to give an accurate opinion, but nonetheless seem unqualified in training or intellect. This expert would have a high genuineness score, but a low skill score. The point is that both sorts of considerations (and
scores are added together to give an overall credibility score. A factor analysis confirmed that credibility can be usefully conceptualized as composed of these two factors.50

Both constructs were significantly affected depending on whether a blind expert was used (skill: \(F(2, 272) = 13.93, p < .001\); genuineness: \(F(2, 272) = 37.77, p < .001\)).

However, as one might expect, the blind procedure had a greater impact on genuineness.51 When an expert is blinded, the jury perceives greater skill and genuineness compared to the counterfactual situation where neither expert is blinded.

But for the non-blinded expert in a blinded expert condition, the jury perceives the non-blinded expert as having lower genuineness but the same skill level as in the counterfactual condition.

<table>
<thead>
<tr>
<th></th>
<th>Plaintiff’s Expert</th>
<th>Defendant’s Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(component)</td>
<td>(component)</td>
</tr>
<tr>
<td>Trustworthy</td>
<td>.956</td>
<td>.907</td>
</tr>
<tr>
<td>Honest</td>
<td>.956</td>
<td>.922</td>
</tr>
<tr>
<td>Fair</td>
<td>.833</td>
<td>.934</td>
</tr>
<tr>
<td>Clear</td>
<td>.923</td>
<td>.324</td>
</tr>
<tr>
<td>Logical</td>
<td>.688</td>
<td>.389</td>
</tr>
<tr>
<td>Knowledgeable</td>
<td>.898</td>
<td>.969</td>
</tr>
</tbody>
</table>

50 Principal component analysis converged after five iterations. Factor loadings were as follows:

51 Planned contrasts revealed that perceptions of the plaintiff’s expert’s skill, relative to the Neither BE condition \((M = 3.23, SD = 0.62)\), significantly increased whenever he was the only blind expert \((M = 3.51, SD = 0.05)\), \(t(272) = 3.07, p = .002, r = .18\). When the defendant’s expert was the only blind expert, however, perceptions of skill did not significantly change \((M = 3.10, SD = 0.53)\), \(t(272) = 1.36, p = .18, r = .08\). Assessments of genuineness also significantly increased in the Plaintiff BE condition \((M = 3.43, SD = 0.67)\), relative to the Neither BE condition \((M = 3.16, SD = 0.65)\), \(t(272) = 2.48, p = .014, r = .15\). When the defendant’s expert was the only blind expert, perceptions of the plaintiff’s expert’s genuineness significantly decreased \((M = 2.63, SD = 0.07)\), \(t(272) = 4.85, p < .001, r = .28\).
Open-ended responses

After rendering their verdicts (but before being prompted with expert credibility questionnaires), participants were required to provide a sentence or two explanation of their decisions. Participants were also allowed to provide any other comments at the very end of the study. All of these responses were independently coded by two research assistants, with coding discrepancies resolved by the principal investigators.

The participants’ open-ended responses were largely driven by the facts in the case, as one would expect. For example, one juror said that: “While it is understandable to think now that the Defendant should have done the MRI that would have identified the problem, it [is] also understandable to think that he would be conservative about his diagnosis and choose instead to continue observation instead of charging his patient for a test that may not be necessary.” Others invoked policy considerations and mentioned particular facts, for example: “These kinds of lawsuits drive up medical costs for everyone, [and] the doctor met the standard of care based on the information he received from the patient and the tests in his office. Plus the patient did not go back sooner than 3 months.” These themes about the standard of care were raised by the experts and attorneys throughout the trial.

Of those who were in blind expert conditions, 43 (20%) explicitly mentioned the blind procedure or the blind witness favorably (including those referring to the expert being picked from a “pool” or referring to the other expert as “hand-picked” or “cherry-picked”). Eleven more participants (5%) referred to the “bias” of the unblind expert,
without referring to him as such, which totals 25% of the subjects who invoked blinding or witness bias as a basis for their decision. For example, one wrote that “the plaintiff’s expert witness, being a blind witness, had no reason to be swayed.” Another argued the facts of the case and then said, “plus, having a blind expert does make me feel more confident in Dr. Dennis’ [the defendant’s] decisions.” Some seemed to think that this was a determinative factor, writing that “it’s clear to me that the plaintiff’s doctor has been purchased and will be biased to say whatever he needs in order to get his money. His opinion is worth little to nothing to me, whereas the integrity of the blind witness is inherently honest.” Others mentioned the blind procedure only to say that it was “not the determining factor in my decision” or that “it did not sway me either way.”

Discussion and limitations

The primary question for this study was whether an expert’s testimony would be perceived as more credible, and thereby more strongly influence juror verdicts, if he or she was blinded. If so, then litigants would have an incentive to use the procedure, and it would actually have some chance of improving litigation, to better and more legitimately serve its deterrence, compensation, and punishment functions.

The results indicate that using a blind expert provides a significant and meaningful benefit in terms of achieving a verdict favorable for one’s client. The plaintiff in this case was at a disadvantage when neither expert was blinded – such as in a typical trial – winning only 45.9% of the verdicts. But when the plaintiff used a blind expert, he was able to overcome the initial disadvantage and secure a majority of the verdicts (62.3%) from the jurors. A similar movement was observed in the other
direction when the defendant’s expert was blinded. Collectively, this amounted to a
doubling of the odds of winning, which may be worth hundreds of thousands of dollars
or more, depending on the damages at stake in the case (see Figure 1). And it is worth
emphasizing that the actors and substantive testimony of the case were identical across
conditions, and that participants were randomly assigned to experimental conditions,
including a control. Thus, this finding allows a strong inference of causality.

In the real world, the impact of blinding may be even greater if litigants use blind
opinions as screening devices for determining which cases to litigate and which to drop
(or settle quickly). If blind experts are more objective and accurate, then the cases upon
which they provide favorable opinions will likely have other evidentiary features that
cohere with the expert testimony, thereby strengthening the persuasiveness of the
expert and otherwise enhancing the chance of winning. This “selection bias” towards
meritorious cases was controlled for in the present experiment, since the facts and
substantive testimony were the same across conditions, and blind experts were
assigned to either side for experimental purposes. A litigant’s strategy of getting early
blind opinions and then litigating favorable cases may be even more lucrative than the
$100,000 to $150,000 benefit calculated above (see Figure 1).

Thus, it may turn out that the blind expert creates a new equilibrium, wherein
each party invests in purchasing a blinded expert opinion. In bona fide fields of scientific
knowledge, one would expect that blinded experts will often agree with each other,
since their opinions will not be polarized by the influence of litigants. If that is true, and
if litigants are effectively limited to a single blinded opinion (Robertson 2010, p. 212),
then in most cases, a single blinded expert will actually appear for trial, since the other blinded expert’s opinion will be hidden in attorney-work product protections (id., at p.211). Both sides will bring blinded experts to trial only in the rare cases where the two blinded experts disagree, and the disagreement happens to be distributed in a way that is favorable to each sponsor (id., at 217). We would hypothesize that when jurors are presented with two blinded experts, they decide the case similarly as if there are two unblinded experts. This rare outcome would thus impose costs on the litigants, without providing a relative advantage. Future studies could explore this possibility. Rational litigants would need to account for such a potentially wasteful outcome in evaluating the net benefits of the procedure. (See id., at 256-257, providing an equation for that purpose, and showing that on a wide range of assumptions, the tactic remains a rational strategy).

This experiment did not include a deliberation phase or allow jurors to vote in groups to render collective judgments, which would be the ideal dependent variable of interest. However, the six-level juror certainty data provides a useful proxy. One can imagine a sample of 12 jurors being drawn from the observed jurors in each group, and one might further hypothesize that jurors who have very strong personal verdicts will have a greater impact on the collective jury decisions. In the condition with no blind expert, with substantial portions across all six levels of the Likert-scale, any given sample could be skewed either way; as such, both parties face substantial risk of unfavorable ultimate decisions. The blind expert conditions, in contrast, pushed jurors to higher levels of certainty in accordance with whichever side brought the blind expert. In these
conditions, the party with the blind expert is likely to get very favorable initial votes
from a sample of 12 jurors, and also has some assurance that the outcome of
deliberations will reflect this initial distribution. Thus, the predictability of jury decisions
may be improved, which would be an important benefit for litigators (Jacobson et al,
2011).

We predicted that the effects on verdict would be mediated by perceptions of
credibility. To explore this possibility, participants were asked to rate both expert
witnesses on six factors from which a composite score of overall credibility was
constructed. As hypothesized, credibility was significantly enhanced when the expert
was described as a blind expert, and this was true for both the plaintiff’s expert (6.39 to
6.94) and defendant’s expert (5.96 to 6.28).

Not only did blinding an expert boost the credibility of that expert, but it also
*harmed* the credibility of the *opposing* expert. Apparently participants perform a
contrast between the two experts to determine their relative credibility. Thus, when
the plaintiff retained the blind expert, the defendant’s expert suffered a reduction in
credibility from 5.96 to 5.35; likewise, when the defendant retained the blind expert,
the plaintiff’s expert suffered a reduction in credibility from 6.39 to 5.73. Closer
examination of the skill and genuineness subdivisions reveals that this effect is driven by
altering the relative genuineness of the unblind expert, which reduced from 3.16 to 2.63
for the plaintiff’s expert and from 2.87 to 2.44 for the defendant’s expert. Assessments
of skill, on the other hand, although being lower, were not significantly different.
Litigants may exploit this effect by using their own blind expert to paint the adversary’s
4. Effect of blinded experts on juror verdicts

Expert as an unreliable “hired gun,” notwithstanding whatever impressive credentials or technical abilities he or she may possess.

This loss in credibility may have larger consequences for jury trials, where litigants speak through their witnesses and their attorneys. Trial attorneys are advised that, “Your case will be presented largely through witnesses. If the witnesses are not credible, you have no chance of establishing your own credibility” (Easton, 1998, p. 8). If jurors perceive that one party is taking special precautions to ensure that the evidence is fair and unbiased, that party may also be able to present other evidence in a more favorable light, getting the benefit of the doubt from jurors. Likewise, closing arguments from such attorneys may be more persuasive. And blinding may fit into the larger narrative of a trial, showing which side is taking appropriate precautions and acting reasonably, versus which side is acting negligently or even recklessly in the light of known problems, such as witness bias (Ball & Keenan, 2009).

These conjectures about spillover effects are fodder for future research. However, the observed pain and suffering awards in this study are suggestive of such a spillover effect. Using a blind expert may have an impact on the amount of money that a juror deems appropriate, an effect that becomes significant when combined with the differences in win rates.

Jury instructions were included in two conditions under the hypothesis that, if the judge acknowledged the blinding process as a specific factor for considering the credibility of the experts, especially immediately before asking the jurors to decide the case, then the effects of the blind expert would be more substantial. This hypothesis
was not supported. If anything, there was a counteractive effect, with the benefit of the blind expert on verdict outcome being slightly (but statistically insignificantly) reduced whenever instructions were included. Judicial instructions are notorious for being poorly understood by jurors, though some prior research has documented that they can have a significant effect on outcomes when understood (see generally, Lieberman & Sales, 1999). The given instruction was not particularly strong or emphatic in its support of blinding, concluding merely that, “you may find the blind expert more credible and you may decide to give the testimony of that expert additional weight. However, the ultimate decision about the credibility of witnesses is yours” (emphasis added). Even though it comes just moments before the juror decides the case, this particular wording may have been ambiguous to a juror, trying to determine whether to give more weight to the blind expert. If a judge were to provide a more ringing endorsement of the concept, it may drive jurors further towards the blinded expert. This intervention would, however, also risk an appellate court reversing for the trial court invading the province of the jury.

The important finding is that blinding drives a significant change in outcomes regardless of jury instructions. Finding that judicial intervention is unnecessary adds further credence to the litigant-driven model of blind expertise, wherein self-interested actors would use the procedure to improve the values of their own cases.

This experiment suffered from several limitations. First, the stimulus was condensed and artificial. Although we consulted with practicing physicians and experts on trial practice to design the 35-minute core stimulus video, it was not a real trial, and
indeed was much shorter than a real medical malpractice trial, which can proceed for days or even weeks. The impact of blinding may be smaller in the context of a full trial, if it is swamped by other evidence and argument. Relatedly, we tested a particular sort of medical malpractice case as a stimulus. Blind expertise may have a different impact on other sorts of medical malpractice cases, or even other sorts of civil or criminal litigation.\textsuperscript{52}

Second, it is likely that, if the trial stimulus had provided a more robust discussion of the pain and suffering issue, including requests by the attorney for a specific dollar amount, jurors may have anchored on that reference point. We may have thereby been able to reduce the amount of variance in our data.

Third, the portions of the video stimuli that discuss the blinding procedure were also truncated. For example, a complete discussion of the blind procedure would have included some explanation about whether the litigant could draw multiple times until she got a favorable blind opinion, and then only disclose that one opinion to the factfinder, thereby exploiting a concealed selection bias. In conception, this is not a problem, because there is a waiver of attorney work product protections that would force such a disclosure (Robertson, 2010), thereby functionally limiting litigants to a single blind expert opinion on a given question. We thus omitted discussion of this dynamic for fear of needlessly complicating the jury’s analysis. However, some of the

\textsuperscript{52} Even though the causal impact in this case was observed in both directions, and in about the same size, it is possible that there will be ceiling and floor effects in other cases, where the baseline evidence is extremely lopsided for one party or the other, and the presence of a blind expert may be insignificant. However, it is doubtful that such highly skewed cases would be tried to a jury rather than settled by rational parties. See Vidmar (1995), p. 175.
jurors’ open-ended responses raised this possibility of iterative selection. For example, one wrote that, “I suspect that the prosecution ‘randomly’ selected several blind experts until they found one with the same opinion as the plaintiff.” Another asked, “My question is[,] how many blind witnesses can you legally hire? Did he hire three others who disagreed and cherry-picked the one who saw it his way?” Another asked about multiple selection and said, “the unclearnness on this point made me wonder if the emphasis on cherry-picking of the other expert was really such a big deal.” This uncertainty apparently reduced the effectiveness of the blinding intervention for some respondents. Future experiments (and actual litigators) should clarify and settle this issue for jurors, and thereby may find an even larger effect for the blind procedure.

Fourth, the study was conducted with human subjects online. Prior research has shown that, “the population of Mechanical Turk is at least as representative of the U.S. population as traditional subject pools” (Paolacci, Chandler & Ipeirotis, 2010, p. 411). Known experimental results have been replicated using the Mturk population (Berinsky, Huber & Lorenz, forthcoming). Still, it is likely that Mturkers may be more easily distracted from the trial compared to real jurors, and may even provide junk responses. Such problems would increase noise in the data. In the real world, the effect of blinded experts may be even greater.

---

53 We included a quiz to check whether they watched the video, but decided not to utilize it to exclude respondents, because we had no baseline for how much real jurors pay attention and thus no way to set a threshold for satisfaction. We were concerned that screening out those who were not paying attention might unduly favor our hypotheses tests.
A fifth and related limitation is that our sample was somewhat younger, more female, and more educated than a median American juror, though our randomization succeeded in distributing these differences across conditions. Our regression analyses do not suggest that these differences were driving our results, though future studies should explore the possibility that more educated people better appreciate the blinding concept.

A sixth limitation is that we did not provide the jury with an opportunity to deliberate. The literature on the effect of deliberation is complex, but some research has shown that deliberation can increase polarization amongst jurors, causing more extreme outcomes (Schkade, Sunstein & Kahneman, 2002). If that dynamic applied here, then the effect of blinding may be even greater. Still, further experiments are underway to replicate these findings and investigate the causal mechanisms in more realistic settings.

Seventh, we used a sample size of 275 subjects to populate five conditions in a factorial design, and used a traditional .05 threshold for statistical significance. We conducted (and reported) multiple statistical tests in our dataset, which increased the risk of false positives.

Notwithstanding these limitations, as a randomized controlled trial with 275 jury-eligible adults from a national sample, using real doctors as experts in a videotaped stimulus with proper jury instructions from a judge, the present experiment has indicia of reliability (Bornstein, 1999). The main effects were observed on multiple dependent
variables including verdict, certainty, expert credibility, and verdict size, which together tell a coherent causal story.

**Implications and conclusions**

Importantly, the stimulus video was designed with a right answer to the question of whether the doctor met the standard of care of the medical profession based on a national practice guidelines (Chou, 2007). Even though we did not provide this information to jurors, it provides a point of reference for analyzing jury performance. In the control condition that simulates real-world trials with counterpoised unblinded experts, most jurors (54%) failed to reach that conclusion, on the basis of testimony of two well-qualified experts, who attempted to explain the standard of care. Jurors did worse than coin-flipping, an important and disappointing finding for the debate over jury competence in the status quo (see e.g., Ivkovic & Hans, 2003; Hans & Eisenberg, 2011).

This finding is strikingly similar to that reported in the subset of real-world medical malpractice trials, where scholarly reviewers independently determine that the truth favors the plaintiff, and yet the defendant won nonetheless. For example, Studdert & Mello (2007) found that in such cases, defendants won 57% of the time. While disappointing as a matter of policy, this finding suggests that our experimental stimulus may have ecological validity as a typical medical malpractice case, where juries often get it wrong.

Policymakers and advocates use such findings to argue for limits on the role of jurors in the American legal system, and advance proposals to instead relegate the
factfinding function to other institutions, such as judges, specialized health courts, or ex ante regulators. Still, somebody has to decide these sorts of cases, and “in complicated fields like DNA, epidemiology, or chemistry, judges are also laypersons.” (Vidmar & Diamond, 2001, p.1169). In theory, specialist courts or expert regulators could have more competence, but would suffer from other problems such as selection bias or regulatory capture. Such reforms also face difficult questions of political economy.

One might argue that jurors are simply doing what they are told in the epistemic situation they are given. It may be that, when jurors are given only two counterpoised experts – who are likely biased by their sponsors selection, affiliation and compensation – jurors are left near epistemic equipoise. The judge instructs them to return a defense verdict, given the preponderance of evidence standard, and they comply. In this light, jurors may be highly-functioning components of a dysfunctional system. Rather than changing the decision maker, it may be prudent to instead provide factfinders with better epistemic signals. Blinded experts may fit that bill.

The foregoing data shows that blind expertise more than doubles a litigant’s odds of winning, an effect much larger than we hypothesized, and likely larger than litigants would otherwise assume. If litigants have so far assumed that the effect would

\[\text{\textsuperscript{54}}\]

See e.g., Huber (1985) and Struve (2004), discussing such proposals. For example, in the Fair and Reliable Medical Justice Act, S. 1337, 109th Cong. §3(d)(4)(b) (2005), senators proposed the creation of health courts presided over by judges “with health care expertise,” but who only need to “meet applicable State standards for judges.” Such judges would likely need to rely on expert witnesses, which just replicates the problem of bias.
be little or nothing, that may explain why the blind procedure has not already been utilized in litigation.

The economic findings displayed in Figure 1 – that the presence of a blind expert may be worth hundreds of thousands of dollars to litigants – suggest that there is a very large upside to this strategy. Of course, the vast majority of cases are never tried, but the amounts of settlements are largely based on predictions about potential trial outcomes that will ensue if settlement negotiations fail (Priest & Klein, 1984). In addition to this benefit, a blind expert may also allow litigants to efficiently screen and select cases, thereby reducing litigation costs by causing earlier settlements – dynamics not studied here. A blind expert may also carry extra weight with the trial and appellate judges, thus improving a litigant’s odds of prevailing on Daubert motions, summary judgment motions, and post-trial review of verdicts. It certainly would not hurt.

Litigants will weigh these advantages against the cost of paying in advance for an opinion that may or may not turn out to be helpful. Traditional attorney work product protections allow that a litigant can procure a blind expert opinion without risk of it harming her case. She can discard the opinion if she finds it unhelpful, or supplement the opinion with a traditional expert witness with whom she can exert greater control (a strategy not tested in the present experiment). Thus, on net, the blind procedure seems to be a rational strategy for litigants – they will pay the cost of such a blinded opinion in order to potentially double their odds of winning, and thus improve the expected value of their settlements.
Whether blinded experts will improve the process and outcomes of litigation is another question. Of course, litigant-induced biases are not the only problems with expert testimony. But arguably, to the extent that the litigant-induced biases of unblinded experts degrade the reliability of their testimony, these biases undermine the ability of factfinders to do their jobs. The blind procedure preserves the best parts of the adversarial system, including robust cross-examination, while avoiding the problems of litigant-induced bias. If blinded experts are more reliable guides for the factfinder, and jurors disproportionately rely upon their testimony, then blinded experts will improve the litigation process, making it more legitimate and improving the quality of its compensation, deterrence, and punishment signals.

References


4. Effect of blinded experts on juror verdicts


4. Effect of blinded experts on juror verdicts


CHAPTER 5: USING CONTINUOUS RESPONSE MEASUREMENT (CRM) IN JURY RESEARCH

Author Note

This is work in collaboration with Christopher Robertson. The authors thank Guy Hagen and Watchfire Analytics for providing access to continuous response measurement technology, and Jennifer Cook for excellent assistance on the literature review.
Abstract

Continuous response measurement (CRM) is a method for understanding how subjects make decisions when responding to a stimulus over time. Unlike analyses of outcome data (e.g., jury verdicts), CRM allows disaggregation of a complex stimulus and facilitates the exploration of potential psychological mechanisms. Media and jury consultants now routinely use the tool, but uses seem constrained to qualitative eyeballing of trend charts. In the scholarly literature, CRM has received little attention. The literature on use of CRM with juries is particularly scant. With online video hosting technologies, integrated CRM tools, and online subject pools, CRM is now much more widely available and accessible for research purposes. In this paper, we review the potential of CRM as a research tool and present our own experiment using CRM in a mock trial, replicating and providing new insight into a prior experiment that used only outcomes data. We present several novel methods for interpreting the resulting data, which is voluminous, tends to have high variance, and is abnormally distributed. We discuss the limits of causal inference using this sort of data.

Key words: continuous response measurement; jury research; decision making; introspection; think aloud; Bollinger bands
Background

Continuous response measurement (CRM) refers to the collection of the same variable repeatedly over time. If used properly, it can provide a valuable window into how a person is processing information moment-by-moment, which in turn can empower a more nuanced understanding of how opinions and emotions evolve and, ultimately, emerge as a behavioral judgment or decision. There are many options for how to implement CRM and many more options for how to analyze its results. In this paper, we discuss the history of CRM generally and its use (or rather lack thereof) in jury research specifically. We then develop and offer refined techniques of CRM, notably regarding how to analyze its results. A mock civil jury trial is examined with our CRM approach in order to demonstrate its value in affording insights above and beyond traditional outcome variables, such as final verdict or self-reported rationales after the fact. Our hope is to spur additional research using CRM, especially the use of CRM coupled with controlled experimental methods—a promising frontier that has received virtually no attention to date.

History of CRM

The first recorded uses of continuous response measurement appear in the 1930s (Levy, 1982; see generally Millard, 1992). Paul Lazarsfeld, an Austrian sociologist, used it to study the emotional experiences associated with listening to music. He had subjects flip, at each beat of a metronome, through the pages of a day-by-day desk calendar, marking at each moment whether they liked or disliked the music. Frank Stanton later recruited Lazarsfeld to join his Princeton Radio Research Project, which
sought to understand the psychological aspects of radio listening. Drawing on his experience with polygraph recording, Stanton soon developed a handset with electronic recording equipment capable of continuously capturing audience reactions to a radio broadcast. The resultant device – the Lazarsfeld-Stanton Program Analyzer – consisted of two cylinders (“like a jump-rope handle”), each with a button (either green or red). The user would hold one cylinder in each hand, with a button under each thumb. He or she could press the red button to indicate disliking or the green button to indicate liking, or not press anything when feeling indifferent. (The first device actually had only one button to indicate liking, but subjects quickly complained about the inability to express dislike as well).

The Lazarsfeld-Stanton Program Analyzer was subsequently used for a wide variety of communication studies, for both radio and television. The device itself also quickly evolved and spawned competitors. There were 15 different variants of the machine by the 1950s (Levy, 1982), and today there are countless devices, and firms offering services associated with those devices.

Designers early on appreciated, at least in broad contours, how the response modality could affect the interpretation of results. The advertising agency McCann-Erickson, for example, discovered that many users of the Lazarsfeld-Stanton Program Analyzer would often possess a predominately favorable or unfavorable opinion, yet not press any button. It was a threshold issue: respondents were not pressing a button at the slightest inclination of a valenced opinion, but were instead setting higher thresholds that must first be passed in order to merit thumb action. Multi-point scales
were introduced. McCann-Erickson simply added options for mildly disliking and mildly liking. The firm Audience Research, Inc. used a “Hopkins Televote Machine” with a 5-point scale, spanning “very dull,” “dull,” “neutral,” “liked,” or “very much liked,” to rate movies for the motion picture industry,

Researchers also developed different physical mechanisms for indicating responses. There are push button devices, such as a 10-button handset, whether the numbers might be organized along an ordered spectrum or have categorical meaning. There are dial devices, with knobs that could be turned clockwise or counter-clockwise. There are slider devices, where an indicator of some sort could be positioned along a spectrum to indicate degree of affinity to one of two poles. And especially more recently, there are computer interfaces, where keyboard or mouse input is given, which in turn might be mapped to one of the three prior mechanism (e.g., a mouse might be used to move an onscreen slider).

CRM is used today in select applied and academic settings. Much of its use is in communication studies, such as advertising or political campaign and debate, but by no means all of it. Lazarsfeld would, for example, be pleased to see research still grappling with emotional responses to music (e.g., Schubert, 2001). And most people have been at least exposed to some form of CRM. Most notably, during the 2004 U.S. presidential debates, two ABC News programs broadcast the debate between incumbent George W. Bush and Senator John Kerry with a live graphic of CRM from a group of potential voters plastered at the bottom of the screen—apparently the first use of CRM during a live presidential debate (Weaver III, Huck, & Brosius, 2009).
In sharp contrast, CRM is virtually never used in published jury research. An exhaustive search revealed exactly two published studies on target (Hughes & Hsiao, 1985; Kette & Brandstätter, 1990). Although innovative, neither used the particular methodological approaches we recommend here, nor was there any attempt to link the methodology with a controlled experimental design to test its reliability with more proximate outcomes.

The first paper, by Hughes and Hsiao (1985), involved three studies. The first cast a small number \( n = 25 \) of community volunteers, just released from actual jury duty, as mock jurors hearing a burglary case. The abbreviated trial elements were delivered by video recording to three groups, of four, nine, and twelve persons. Participants used a handheld box with a response dial to periodically report their “leanings” on a numbered, 7-point scale. Anchors included “for the prosecution or plaintiff” and “for the defendant,” with “1 (or 7) = strongly certain,” “2 (or 6) = medium certainty,” “3 (or 5) = slightly certain,” and “4 = no decision at this point” as the intermittent tick marks. For whatever reason, users did not actually continuously rate their leaning. They were instead prompted at preordained moments to make a judgment, in particular after: prosecution opening; prosecution case; defense case; closing arguments; judge’s instruction; and deliberation. Analyzes proceeded by describing the counts in favor of a guilty verdict. For example, “[t]he opening statement of the prosecutor convinced nineteen of the jurors that the defendant was guilty to some degree,” or that “[t]he judge’s instructions altered two jurors’ opinions to the prosecution, two to neutral, and two to the defense” (Hughes & Hsiao, 1985, p. 55).
A second and third study apparently used the same technique with, respectively, two 12-person mock jurors hearing a products liability case and a non-descript class of ten law students hearing a sexual assault case. Bizarrely, the authors fail to actually discuss any results for either study, and instead just assert that they “can report how much each trial altered the jurors’ minds and the average change for the jury, enabling the attorney to isolate the event which locked a juror into one side or the other”—then abruptly proceed to a general discussion of jury selection strategies (Hughes & Hsiao, 1985, pp. 57–58). Following a footnote trial to downstream appendices, one finds tables of the raw data for each juror. An adjacent column computes the average rating at each preordained assessment moment.

The second paper, by Kette and Brandstätter (1990), involved 37 Austrian law students shadowing an attempted murder trial (i.e., they sat in the courtroom audience and listened to the actual arguments). The trial lasted one day, with the opening morning dedicated to evidence presentation (closing and judicial instructions were in the afternoon). Response ratings were collected by paper-and-pencil, on a 10 cm graphic rating scale, from “guilty, very sure” to “not guilty, very sure.” The researchers described their instructions to the mock jurors as follows (emphasis in original): “The subjects were to record any change, even the slightest one, of perceived guilt any time such a change occurred. In addition, the exact time, the speaker[,] and a key-word were to be recorded” (Kette & Brandstätter, 1990, pp. 20–21). Assuming respondents comply with such instruction, this would allow analyses at a very high temporal resolution (to the minute, if not less).
Kette and Brandstätter split the 4-hour evidence presentation into eighty 3-minute intervals and then applied a two-step analytic approach. In the first step, eight independent law students rated each 3-minute interval in terms of five factors: legal relevance, concreteness, emotionality, “stimulation of compassion with the defendant,” and attribution of responsibility. The second step involved autoregressive integrated moving average (ARIMA) modelling. These models are complicated to properly fit and even more difficult to intuitively interpret—and thus of use only to those with econometric training. ARIMA is also usually going to be inappropriate for modeling CRM. The reason is that ARIMA models assume constant variance over time. This is unlikely to true in most jury research situations, if for no other reason than that juror ratings tend to facture into two groups, one in favor of each litigant. (This will be evident in our example below.) On the other hand, if the verdict is obvious, variance would tend to shrink as all jurors converge on the same answer. Kette and Brandstätter pressed forward and, using their fitted ARIMA model, identified correlations between all of their rated factors, minus emotionality, and respondent rating changes. The final model included coefficients for each of the estimated correlations, although again the interpretive enterprise here is something of a herculean task.

Potential for CRM

Why is CRM so underutilized in jury research? One limitation, at least until recently, was perhaps logistical. The techniques of collecting continuous measurements were expensive and cumbersome, and the physical equipment was tethered to a particular location. Modern computers, especially when coupled with internet video
hosting and online subject pools, now render this set of concerns obsolete. Nowadays a researcher can stream the video of a trial online and have respondents, anytime and anywhere, engage in a CRM task via web application. Even with local respondents, the equipment concerns are now obsolete. The vast majority of Americans now have sophisticated computing equipment (smartphones) in their pockets, which can become continuous feedback devices with the download of a simple app (see e.g., Asymco, 2014).

Yet equipment is a partial explanation. Computerized CRM has been around for years. And at least jury consultants seem to be picking up the scent of opportunity: a number of for-profit consulting firms advertise continuous feedback services (as a quick skim of Google results for “jury research continuous feedback” reveals).

The second limitation seems to be knowing what to do with the resultant data, which is voluminous and unwieldy. (See Figure 1 below, plotting individual data from 95 respondents, for a tantalizing preview). How can the researcher find order in this chaos, and devise rigorous approaches for detecting meaningful CRM results? The Hughes and Hsiao (1985) and Kette and Brandstätter (1990) approaches illustrate opposite spectrums of the challenge here. On the one hand, Hughes and Hsiao’s approach was too simplistic, essentially picking out any movement in scores and inferring it was significant (both statistically and substantively) and due to the immediately associated trial content. This approach may be most similar to the ways that CRM is currently used in applied settings. On the other hand, Kette and Brandstätter valiantly tried to force econometric time-series modelling techniques into the courtroom, yet their approach
was needlessly complicated, and still lacked any validation with regard to the real outcomes of interest (jury verdicts). It was laden with (problematic) assumptions and emitted results that were exceedingly difficult to evaluate in a meaningful way. We propose analytic approaches below that are both rigorous, and relatively more intuitive to understand (including transparency in its limitations).

Figure 1. Plot of CRM from 95 respondents.

As the logistic and analytic hurdles are overcome, CRM has the potential to empower a more nuanced understanding of how evaluations shift in response to discrete pieces of information (e.g., trial evidence), evolve over time as information is integrated (or not), and, ultimately, emerge as an overall judgment (e.g., a verdict). This within-subject variation over the course of a stimulus can generate hypotheses, which can then be more rigorously tested in randomized experiments, isolating potential
5. Using CRM in jury research

questions of interest. CRM can provide a window into which parts of a trial—which pieces of evidence—are tipping points in a person’s evaluation of the case. And it can be split by covariates and experimental independent variables, to examine differential CRM trends across subjects.

To appreciate the potential here, first consider what CRM is not: CRM is not the same as a post-decision explanation of why a person decided one way or another.

Researchers of all stripes have long asked subjects, after the fact, why they decided as they did. This includes prompts to think back to particular pieces of evidence presented at trial, and to assess the extent to which those evidentiary pieces did or did not weigh into the final verdict. Yet this sort of retrospective reason-giving is notoriously prone to error and bias (see Nisbett & Wilson, 1977, for a famous review). One important reason is that subjects, when asked to recall their thought processes, often do not draw upon memory of specific events at all, but rather make inferences about the origins of their behavior, almost as if they themselves were a third party observer. In the language of psychologists:

When reporting on the effects of stimuli, people may not interrogate a memory of the cognitive processes that operated on the stimuli; instead, they may base their reports on implicit, a priori theories about the causal connection between stimulus and response (Nisbett & Wilson, 1977, p. 233).

CRM emphatically does not entail this sort of retrospective self-report.

CRM, in sharp contrast to retrospective self-reports, involves a concurrent report of mental states. In this respect CRM is similar to “think aloud” protocols (see generally Ericsson & Simon, 1993). Anders Ericsson and the Nobel-prize winning, polymath Herbert Simon provided an important (if underappreciated) review of how verbal
reports can, if solicited via “think aloud” protocols, provide reliable insights into cognitive processes without suffering the flaws reviewed by Nisbett and Wilson (1977). From a cognitive perspective, the key is having subjects report the contents currently stored in short term memory (STM)—roughly, the stuff of moment-by-moment conscious thought—rather than trying to retrieve information about past thoughts from long term memory (LTM). Neither type of report is a direct read-out of cognitive processes, of course, but concurrent reports are less prone to error than retrospective reports because they do not require the additional processing steps involved in retrieving information from LTM.

**Method**

Our aim here is to investigate the reliability and utility of CRM. To this end we use CRM data from an in-person randomized jury experiment conducted in Tucson, AZ. The experiment tested aspects of an institutional solution to expert witness bias known as “blind expertise,” which is designed to improve the credibility of expert witnesses and, ultimately, improve the accuracy of juries tasked with resolving technical questions. This notion of “blinded” experts has been developed elsewhere (Robertson, 2010), and this experiment allows replication and extension of earlier findings on its use in civil trials (Robertson & Yokum, 2012). Yokum and Robertson (in prep) explain the experimental manipulations in depth and discuss results as pertains to the blinded expert concept. Here we focus on CRM as a research method, investigating its reliability and utility for analyzing the core case, as well as the unique opportunities that emerge when coupling CRM with an experimental design.
Participants

Ninety-five jury eligible adults from the metropolitan area served as mock jurors. The sample roughly mirrored the population characterized by the U.S. Census in terms of education, gender, age, and race (but with a tilt toward being younger and more educated).

Subjects were also scored on the cultural cognition scale (see Kahan, 2012, regarding scale development). Cultural cognition is essentially a sub-species of motivated reasoning, in which the way people psychologically process information (and thus perceive facts) is influenced by their already existing value commitments (see Kahan, 2012, for a review of cultural cognition; see Kunda, 1990, for a general theory of motivated reasoning). The novel cultural cognition hook is its associated scale, which scores people along two value dimensions: hierarchism-egalitarianism and individualism-communitarianism. These concepts build off the “group” and “grid” typology of cultural worldviews developed by Douglas and Wildavsky (1983). Kahan summarizes the basic notions, mapping them onto cultural cognition terminology, as follows (emphasis added):

A “low group” worldview coheres with an individualistic social order, in which individuals are expected to secure their own needs without collective assistance, and in which individual interests enjoy immunity from regulation aimed at securing collective interests. A “high group” worldview, in contrast, supports a solidaristic or communitarian social order, in which collective needs trump individual initiative, and in which society is expected to secure the conditions of individual flourishing. A “high grid” worldview favors a hierarchical society, in which resources, opportunities, duties, rights, political offices and the like are distributed on the basis of conspicuous and largely fixed social characteristics—gender, race, class, lineage. A “low grid” worldview favors an egalitarian society, one that emphatically denies that social characteristics should matter in how
resources, opportunities, duties and the like are distributed (Kahan & Braman, 2006).

Crossing the two dimensions creates a four-quadrant grid, upon which a person can be classified in terms of “hierarchical individualism,” “hierarchical communitarianism,” “egalitarian individualism,” or “egalitarian communitarianism.” Applying the scale, one can then test hypotheses related to whether persons of different cultural cognition backgrounds perceive the same facts differently.

Mock trial procedure

Subjects were seated at private computer stations, where they watched a 35-45 minute mock trial video as stimulus (variable length due to the different experimental conditions), with audio provided over headphones. The video contained the key elements of a real trial: instructions from the judge, opening statements from attorneys for the plaintiff and defendant, examination and cross-examination testimony of two expert witnesses (one for the plaintiff and one for the defendant), closing arguments, and finally closing instructions from the judge.

The trial itself concerned the failure of a primary care physician to diagnose a possible case of lumbar radiculopathy and refer the patient to imaging, which allegedly would have allowed timely surgery and avoidance of the permanent disability that the patient now suffers. The primary dispute concerned whether the physician-defendant met the standard of care when, instead of ordering imaging, he simply instructed the patient to take pain-killers and return if the pain got worse. The litigant arguments were designed to seem relatively balanced. In favor of the plaintiff, for example, were positive tests of two neurological tests for injury, as well as plausible activities (physical
exertions at the gym and garage) within which the injury might have been suffered. The defendant, for his part, advocated for the benefits of conservative medicine—coupled with the safety net instruction to return if the injury worsened—and also insinuated that the plaintiff might be engaged in drug seeking behavior.

**CRM – User Input**

There are multiple ways to collect continuous response data from the user, from handsets with dials or buttons, to mouse or keyboard computer interfaces. Which is best? Here respondents used a computer mouse to slide a ball left or right on a scale positioned at the bottom of the screen, just below the stimulus video (see Figure 2). It included the anchor labels “defendant” on the left and “plaintiff” on the right, and nine horizontal lines as if a 9-point Likert scale. The actual collection was even more precise; the ball could move smoothly throughout the scale (as opposed to snapping onto the vertical line positions), and the precise position was captured in integers ranging from zero to 100.
The physical properties of the collection device, and visual representations thereof, are potentially important for at least two reasons, one common to all surveys and one unique to CRM. The common survey issue is that the answer given by a respondent can be subtly influenced by aspects of scale labelling and response modality. Studies have, for example, demonstrated that responses can fluctuate according to the number of options on a Likert scale (e.g., 7 versus 10 options, as in Preston & Colman, 2000), the range over which the same number of options spread (e.g., 0 to 10 versus -5 to +5, as in Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991), or whether a middle neutral option is provided (Bishop, Tuchfarber, & Oldendick, 1986). These subtleties can be important, but not always or even usually. At any rate it is not unique to CRM, and thus does not merit extended discussion.

We make two quick notes though. First, in our experience, we have seen mock jurors become confused by the labels “plaintiff” and “defendant,” sometimes forgetting
how the legalistic jargon maps onto the personal names of the litigants. Our antidote here was low tech: we simply tapped a piece of paper on the bottom of the monitor, reminding users that the patient Mr. Stevens was the plaintiff and Dr. Dennis the defendant. Second, it is possible that the label ordering could tilt the direction of reasoning. People might, for example, search for reasons in favor of whatever litigant is on the right-most side. This could be important, as several studies have shown that searching for reasons in favor of a choice can lead to different results than searching for reasons against a choice (see e.g., Shafir, Simonson, & Tversky, 1993). So for example, if “Defendant,” is on the right, a juror might parse the trial with a keen eye on reasons for the why defendant should win, and be less likely to notice favorable plaintiff facts, and vice versa had “Plaintiff instead been listed on the right.

The second, CRM-specific issue is whether the collection device draws and keeps attention over time. The concern here is distinguishing a stationary rating as indicating indifference—a lack of opinion change at that moment—rather than the user simply forgetting to indicate his or her opinion change. One can conceive various tactics to combat this problem, with varying levels of intrusion. The earliest tactic, by Lazarsfeld in the 1930s, used a clicking metronome to set the pace with an auditory cue. With a computer, an instructional dialog box could surface at either predetermined intervals or after the user is stationary for a set period of time, to insist on either a change or confirmation of no change. Such tactics have the downside of being intrusive and, therefore, potentially disruptive of the naturally unfolding thought process. A different strategy is tactile. The Televac system from the fifties, for example, entailed a joystick
device that exerted a slight pressure back to the central, neutral position, and thus required the user to exert constant muscular tension to hold a non-zero rating (Millard, 1992, p. 11). Televac users were less likely to absentmindedly sit on a positive rating (although it seems they could still just as easily forget to move off the zero set-point).

Yet another possible tactic, which we used, is visual: place the input device in a salient position. Here the scale was immediately below the screen frame wherein the video was played. A participant looking at the monitor had both the trial stimulus and CRM scale within the visual field. Ultimately, which of these tactics (or other unmentioned ones) possesses the best properties in terms of ensuring continuous user input is an open empirical question. It might not be the visual one we deployed (see e.g., Simons & Rensink, 2005, on “change blindness,” referring to how people sometimes fail to consciously notice even large changes in their visual field).

**CRM – User instructions**

Before beginning the trial video, research assistants gave participants a verbal and visual tutorial on how to use the scale. Respondents were also given written instructions: “Move the ball to indicate the degree to which you think the defendant or the plaintiff should win the case, based on what you have heard so far. Evaluate constantly as you watch the trial.” It was emphasized that they should continuously rate, ensuring to record any change of evaluation by moving the scale ball.

**CRM – Momentary versus cumulative**

Note that a CRM value could be used to track one of two different types of self-report: (1) an evaluation of the particular evidence presented at that moment; or (2) an
evaluation of how the respondent would likely decide the case at that moment, reflecting the accumulation of evidence up until that moment. Call the former “momentary CRM” and the latter “cumulative CRM.” With momentary CRM, respondents would keep the ball at the center of the scale whenever unmoved by a piece of evidence, and only shift to the left or right when the onscreen evidence was causing them, at that second, to shift their opinion further in favor of the defendant or plaintiff, respectively. As soon as the onscreen evidence stopped having any effect on opinion, the ball would be moved back to the neutral, center position. With cumulative CRM, the ball would also shift left or right as opinion swayed toward the defendant or plaintiff; however, rather than snapping back to the central point of the scale, the respondent would simply stop moving the ball—and leave it wherever it was on the scale—whenever the evidence stopped having an effect on opinion.

Figure 3 provides a simple illustration of the difference. Both panels show hypothetical CRM for the same respondent, where his or her opinion shifts 10 points toward the defendant around minute three and then another 10-point shift occurs around minute 15. The left panel shows momentary CRM, while the right panel shows cumulative CRM.
In the simplest theory the two response modes should generate equivalent data, with the researcher being able to infer one graph if provided data on the other. Whether respondents in fact express their opinions so additively and consistently are open empirical questions. Research in judgment and decision making (JDM) has shown that, in a variety of settings, the selected response mode can influence expressed preference, in some cases so dramatically as to cause outright reversals of preference (see generally Slovic, 1995). Such work suggests the match between momentary and cumulative CRM is unlikely to be perfect.

Consider a classic preference reversal experiment by Lichenstein and Slovic (1971). They asked people to indicate a preference between two possible gambles (A or B), but varied whether the expression of that preference was via choice (i.e., pick either A or B) or pricing (i.e., state how much you would pay to play A and pay to play B). For certain carefully matched pairs, subjects would choose A but say they would pay more
to play B (or vice versa)—a clear violation of the notion that people possess coherent preferences. The psychological explanation in this case is that the different response modalities direct attention to different aspects of the decision. Making a choice tends to focus attention on the probabilities associated with the gambles, while pricing the gambles instead focuses attention on the dollar amounts to be won or lost.

But it is difficult to generalize here. The most straightforward conceptual parallel (from the classic gamble example to the CRM task) would be that momentary CRM draws attention to one type of evidentiary piece while cumulative CRM draws attention to a different type. Yet it seems unlikely that the two CRM modalities would draw attention to different types of evidence within the trial itself. Or at least it is difficult to predict such effects from any a priori theoretical framework. There is no reason to think, for example, that one modality or the other would lead to greater juror focus on witness characteristics, or to weighing forensic evidence differently than eyewitness testimony, or to being more likely to ignore a judicial instruction, or . . . —even trying to come up with plausible examples is difficult.

A more promising approach is to consider whether one of the two response modalities maps more naturally onto a juror’s actual cognitive processes—or more particularly, onto those cognitive processes accessible to consciousness (and thus capable of concurrent self-report). To put it loosely, as people parse the trial moment-by-moment, do they feel themselves making evaluations of each discrete piece of information, independent of their global evaluation of the case? In such case momentary CRM maps naturally. Or do they instead experience only the global
evaluation, one that evolves over time as it instantaneously incorporates each new piece of information? In such case cumulative CMR maps naturally.

Research is needed to sort out the above questions. We chose cumulative CRM for our first foray into this territory, although more out of personal preference than any robust theoretical reason—we need more data, before theory driven choice is possible. To the extent we had a theory, it was the intuition that the moment-by-moment experience of evaluating a case reflects an evolving global judgment, not a disconnected feeling of evaluating each piece of information as if no other information had been presented before it.

**Verdict and monetary damages questionnaire**

After viewing the trial stimulus, all participants were asked to render a binary verdict, answering “Yes” or “No” to the question, “Based on the instructions provided by the judge in the video, do you believe that the Plaintiff has proved, by the greater weight of the evidence, that the Defendant committed medical negligence?” Subsequent questions asked for a verdict rating on a 6-point Likert scale, from 1 (clearly not medical negligence) to 6 (clearly medical negligence), and—if finding for the plaintiff—an assignment of monetary damages for pain and suffering.

**Self-reported rationale**

After rending their verdicts and (if applicable) monetary damages awards, participants were also asked to explain the reasoning behind their decisions, via a short written prompt. (“In your own words, please provide a sentence or two explaining your decision on this page.”)
**Linking CRM and trial segments**

To link the continuous data to particular trial segments, we first transcribed all audio from the videos. (There were 301 sentences in the control video and an additional 184—for a total of 485—sentences spliced into the treatment video.) A timestamp was attached to the beginning of each sentence. For example, the plaintiff’s attorney asked his expert witness, “Did the doctor do an examination,” at 12 minutes and 49 seconds into the trial. We also tagged at a more granular level the major segments of the trial, such as the plaintiff’s attorney’s opening statement or the judge’s reading of the jury instructions. We then used the transcript timestamps to cross-reference the continuous-response data. The linked transcript-CRM data can then be used to assess the potential causal impact of each sentence (or segment) on the respondent’s shift in verdict preference at the nearby moment.

**Results**

**Verdict**

Forty percent, 95% CI [.30, .51], of jurors found in favor of the plaintiff overall. (See Yokum & Robertson, in prep, for discussion of the full experiment and results).

**Self-reported rationale**

The juror’s self-reported explanations for their decisions were coded along several dimensions. Codes were developed by have two independent reviewers read through all responses, categorizing the thematic content of responses. This was done iteratively until each felt they had adequately created a code for each theme. Coding
books were then compared, with the reviewers working to create a common coding. A third reviewer was available to reconcile any disagreement. Once the codebook was finalized, the two independent reviewers again read the written entries, this time scoring along the agreed dimensions.

Four coded themes are notable for present purposes. The first is plaintiff responsibility, or rather irresponsibility: the doctor had told the plaintiff to return if the condition worsened, but the plaintiff failed to do so. Forty-nine jurors listed this as a reason for finding against the plaintiff; four persons affirmatively stated this was not a persuasive argument. The second theme was to endorse (or reject) the defendant’s argument that a conservative course of treatment, rather than rushing to take an expensive MRI, was appropriate. Fifteen jurors endorsed this logic in their rationale; seven affirmatively rejected it. A third common theme, mentioned by eight jurors, was to assert that, whatever merits to the plaintiff’s argument, he nonetheless failed to satisfy the burden of proof.

And finally, narcotics: the defendant’s attorney had insinuated drug seeking behavior, that the plaintiff was feigning pain to get a fix. Notably, only three jurors mentioned drug seeking behavior factored in their reasoning against the plaintiff. Four mentioned they did not find this a compelling argument. Make a mental note here. As we shall see, CRM reveals that narcotics had a much more powerful influence than jurors either recalled or were willing to report.
Looking at the CRM data

Responses were sampled at half-second intervals. With 95 participants watching either a 35-minute or a 47-minute video, this resulted in a relatively large dataset, at least for jury researchers: over 546,000 data points. Figure 1 above shows a plot of each juror’s rating over the duration of the trial. The dataset entails very large inter-subject variation (that is, variance in ratings across different subjects) at any specific moment. This increases throughout the course of the trial, as jurors polarized into different ultimate votes. In the treatment condition, for example, the inter-subject standard deviation grew from 9.8 points (on the 100 point scale) at minute 7 (at the end of opening statements) to 26 points at minute 46 (at the end of closing arguments).

Figure 4 charts mean response within one of the experimental groups (the treatment). This is, in other words, an example of a within-subjects chart. We are not making comparisons between groups, but rather want to know when the evidence meaningfully shifts the verdict preference of the same person. But what constitutes a meaningful shift?
Figure 4. Within-group chart of mean CRM over time.

There is a real risk of the researcher reading the tea leaves here, “seeing” changes where he or she expected to see changes a priori, or mining the data to cherry-pick shifts that fit in line with whatever story is trying to be told. This need not, and presumably rarely would be, a deliberately mischievous sifting of results (see generally Simmons, Nelson, & Simonsohn, 2011, discussing biased statistical decision making, and how “[t]his exploratory behavior is not the by-product of malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions and (b) the researcher’s desire to find a statistically significant result”). We tackle this problem in
the next section, but first look at one more example of the data, this time splitting between subjects.

Figure 5 charts mean response split by final verdict. The upper green line (“Yes MedMal”) shows the average ratings for those jurors who ultimately side with the plaintiff on their verdict questionnaire; the blue “No MedMal” line shows those who ultimately rule in favor of the defendant.

Two points of polarization are visually obvious, around minutes 12 and 21. Other fluctuations seem to exist as well, but again it is difficult to confidently tell which shifts reflect meaningful differences as opposed to random error. Around the 36th minute, for example, the No MedMal group seems to respond more strongly (the slope of the upward moving line is steeper), but is this a reliable difference?
5. Using CRM in jury research

Figure 5. Between-group chart of CRM, split by verdict.

**Statistical approach**

What can be done to help the researcher more rigorously decide which CRM segments represent meaningful changes? We want a procedure that ameliorates some of the risks associated with subjectively eyeballing the chart. We use here a statistical decision rule that flags when a trial segment is potentially meaningful, and thus merits further consideration by the researcher.

Our primary dependent variable was the mean score of subjects within the group under consideration. Figure 4, for example, involved the mean score of all subjects in the treatment condition; in Figure 5, all respondents were included, but this
time averages were conditional on final verdict. Note that different statistics could be chosen (e.g., medians or 10% trimmed means).

Note that simple before-and-after measurements—e.g., subtracting the mean CRM at the end of a sentence from the mean CRM at the start of the sentence—could be highly misleading, for a variety of reasons. There is presumably a time lag, for example, between the utterance of a sentence and when that information had been parsed, comprehended, integrated into verdict preference, and ultimately converted into a physical movement of the scale. There are likely contrast effects: the same piece of information might appear to elicit a smaller or larger reaction depending on whether the immediately preceding sentence happened to involve a large, temporary fluctuation.

Nonetheless, the data do show robust inter-temporal consistency. This can be assessed by taking any 60-second time period, and then measuring the standard deviation of subject response scores. The average of all such standard deviation measurements was only 0.95 of a point on the 100 point response scale. The researcher can take advantage of this general consistency within such 60-second time frames to develop a null hypothesis: If a given sentence has no real effect on jurors, then the mean score should not vary more than 1-point compared to the average score of the preceding 60-second period. With such a null hypothesis in mind, the analyst can identify periods that have much larger movements than expected and, with thoughtful caution, infer causation from the preceding portion of the stimulus video.
To implement the above logic, we first calculated a 60-second moving average (60MA). In particular, we calculated 60MA at time $t$, $60\text{MA}(t)$, as \( \frac{\sum_{i=t-n+1}^{t} r_i}{n} \), where $r$ is the mean rating at time $t$, and $n$ is how many seconds to look backwards, in this case 60 seconds. For example, if a respondent provided a rating of 20 for thirty seconds followed by a rating of 40 for another thirty seconds, the 60MA at the end of that minute would be 30. This moving average is used to smooth out random short-term variations and thereby provide a more stable benchmark for assessing relative change. Figure 6 charts mean score (blue line) as well 60MA (thick orange line) over time.
Figure 6. Mean CRM in the treatment group, with 60MA and “Bollinger Bands” (i.e., upper and lower lines computed as 60SD × 2) to aid statistical inference.
What constitutes a meaningful shift from the 60MA? For each period, we also calculated the standard deviation of the preceding 60 seconds of mean scores—that is, a moving standard deviation, or 60SD. We constructed an upper band that was two standard deviations above the 60MA; a lower band was likewise created a two standard deviations below the 60MA. This is a technique also used to analyze stock market trends, where the 60MA ± (60SD×2) bands are also known as “Bollinger bands,” after John Bollinger, who popularized the approach (see generally Bollinger, 2001). These are shown in Figure 6 as the dashed, dark blue lines.

The decision rule is to flag as statistically significant any mean score that exceeds 60MA ± (60SD×2). In terms of Figure 6, this means we look for mean scores (blue line) that exceed the upper or lower Bollinger bands (thin orange lines). See also Figure 7, showing a zoomed in portion of the CRM chart and highlighting in yellow a region identified as statistically significant according to our analytic approach. It asks whether the mean score at each moment would be unlikely given the mean and variance observed over the prior sixty seconds. The bands are created by multiplying by two standard deviations to achieve the Fisherian norm of $p < .05$: assuming the ratings are normally distributed, we would expect 95% of the means to be captured within the Bollinger bands. But there is nothing magical in that decision. One could easily adjust the band width to achieve any desired $p$ threshold, depending on how one wishes to balance the risks of false positives and false negatives. And, as with any statistical decision rule, the intention here to provide disciplined analytic guidance, not install a dogmatic rule (see e.g., Nuzzo, 2014).
Using the bands shown in Figure 6, we identified 494 instances (amounting to 247 seconds or 4.1 minutes of the video) when the mean score was more than two standard deviations above the 60-second moving average. We identified another 535 instances (amounting to 267 seconds or 4.5 minutes) when the mean score was more than two standard deviations below the 60MA. In order to cluster multiple half-seconds into identifiable transcript segments, we imposed the additional requirement that the deviation remain outside of the bands for at least two seconds. This process identified 55 segments of the transcript. It included 24 shifts in favor of the plaintiff, and 31 shifts in favor of the defendant. The Bollinger band filter thus gave us a sense of large moves
relative to the amount of variation observed across rolling one-minute segments of the video.

We also sought to identify the most significant changes in the context of the changes observed in the entire video. For each half-second period, we subtracted the mean score from the 60-second moving average. We then identified the top 5% and bottom 5% of all the half-second periods, which created a threshold of about a 2.7 point move compared to the 60MA, in either direction. This screen, in combination with the Bollinger band screen, identified 560 such periods, which were clustered into 15 consecutive ranges of the transcript (again ignoring any temporary gaps of two seconds or less). The 15 transcript excerpts are listed in Attachment D below. Figure 8 identifies 10 of those shifts on the chart for one of the experimental groups.
Figure 8. Portions of trial analytic approach flagged as causing reliably large shift.

The flagged transcript portions (and resultant chart, such as Figure 8) provide a wealth of data to the researcher or applied litigator. Some of the findings reinforce other outcome measures; for example, plaintiff responsibility was repeatedly mentioned in self-reported rationales as an important factor, and we see a big dip (around minute 21) when the testimony was highlighting that the patient failed to return to the physician’s office. But there are also insights that simply never surface in the non-CRM data.

Consider Figure 9, showing a zoomed in portion of the CRM chart from a different experimental group, and also superimposing the associated dialogue from the trial transcript. This is the portion of the trial wherein the defense attorney insinuates
that the plaintiff was seeking drugs. Recall that virtually no juror mentioned this as a fact. Yet the CRM shows it is clearly damaging, causing two notable dips in verdict preference (recall lower scores indicate a greater likelihood of finding for the defendant).

Figure 9. Zoomed in portion of CRM chart, illustrating segments identified as significant by analytic method.

Litigators, or their jury consultants, could also use such data to fine-tune their case. In this instance, for example, we planted two key pieces of evidence in favor of the plaintiff, namely, that he scored positively on two neurological exams. According to practice guidelines, this should have triggered prescription of an MRI. Yet the CRM data shows that jurors were unmoved by this evidence—ratings never broke outside the
Bollinger bands. This could be a flag to the litigator to revisit how that evidentiary fact was being presented (e.g., perhaps it was confusing, so a different explanation is required). Note that had we directly asked jurors whether they considered the neurological exams important—as jury consultants might very well do during a focus group, for example—it would not at all be surprising to hear affirmative answers. Most people would, upon reflection, probably agree that a key neurological exam should impact their decision, and thus accede to any suggestion that it has in fact done so. The CRM data avoids this post hoc rationale pitfall, and lets the researcher peer more directly into what actually happened.

**CRM split by covariates**

The CRM data become richer still when one begins to look between-subjects. Consider first a split on covariates, in this instance cultural cognition. Figure 10 charts CRM split by which quadrant of the cultural cognition scale a person fits within. Three moments when particularly precipitous drops occur—but, notably, only for hierarchist-communitarians—are demarcated by boxes A, B, and C. All were flagged as significant departures beyond their respective Bollinger bands (not shown).
These three moments entail the defendant attorney’s insinuation that the plaintiff was engaged in drug seeking behavior. At box A the plaintiff’s expert mentions that “[Mr. Stevens] had been on and off some narcotics for the past five years.” At box B, the defendant’s attorney is crossing the plaintiff’s expert, pressing insinuation:

[Defendant’s attorney]: So isn’t it true that a conservative course of pain management is particularly appropriate in these circumstances?

[Plaintiff’s expert witness]: Well I, I think if what you’re getting at is that a person who has taken narcotics before is somehow less reliable in their description of pain, that is an understandable question. However . . .

And then again at box C, yet more on narcotics

[Defendant’s attorney]: And isn’t it actually quite common for patients to, to act on this point, to pretend in order to get more narcotics prescribed?
[Plaintiff’s expert witness]: That certainly has been known to happen.

The use of CRM within-subjects uniquely flagged that the accusation of drug-seeking behavior was important; the coupling of CRM with the covariates further revealed that this effect was driven by a particular sort of person, namely hierarchist-communitarians.

**Discussion**

Continuous response measurement (CRM) offers a rich source of data, and we considered throughout this paper the various ways it can be collected, such as the type of user input tool or how the scale is labeled. There are several issues to work out in terms of how the chosen methodology might or might not impact respondent behavior (e.g., whether cumulative and momentary CRM render identical results). Which exact methodology is best likely depends on the aims of the researcher. We also considered the various ways the CRM data can be analyzed, and we offered a statistical approach based on the 60-minute moving average. Our aim was to place some disciplined rigor over the interpretive enterprise, while also using a method that was relatively straightforward to understand and, thus, transparent for purposes of peer review. We close by further reflecting on the empirical data, and in particular its value above and beyond traditional outcome measures (e.g., verdict or post-hoc rationale), and then more generally on how CRM can be coupled with both within-subjects and between-subjects methodologies.

CRM provides a window into temporal dynamics of a trial that a static questionnaire after the fact simply cannot offer. One important observation here, for
example, is how quickly jurors potentially make up their minds. Note how, in Figure 5, jurors are clustered as early as the 7th minute according to their ultimate verdict. This is the moment immediately after both parties have given the opening statements.

Countless lawyers and trial consultants advise that cases are won and lost by the openings statements (Goleman, 1994); still others opine that such suspicions merely reflects “[a] common myth that pervades the legal world” (Trial Behavior Consulting, n.d.). Our data suggest that opening statements might, in fact, be relatively decisive as some litigators speculate.

Note how other studies have tried asking jurors to recall when they made up their mind, and instead find self-reports of much later decision making. Waters and Hans (2009), for example, asked jurors, “When did you start leaning towards one side,” and observed a mean answer of 4.5 out of a 8-point scale (from 1 = during prosecutor’s opening statement, to 8 = during deliberations). This is a good illustration where CRM and post-hoc reason-giving can diverge. And in this instance, we place our chips on CRM, given its unique advantages in this circumstance. Self-reports such as those solicited by Waters and Hans (2009) suffer at least two serious risks on this question. The first is that people are unlikely to actually recall the exact moment they started to make up their mind. The second is that, even if they do recall, it would likely be an embarrassing concession to have to state publicly, to the researcher, that one is, basically, a close-minded person—someone who makes up their mind before hearing all the evidence. This is a particularly unpleasant admission in the jury context, where one is charged with fairly weighing all the evidence. The CRM data avoid these problems. At
any rate, CRM could be used to further study this issue, and test whether hypotheses such as these bear out.

This illustrates a more general benefit of CRM, namely, that as a form of “think aloud” data, it can capture information that subjects are later unable to accurately recall (or perhaps unwilling to report upon reflection). We observe in our example case a notable instance of this dynamic. Based on self-reported verdict rationales, we would not have identified drug-seeking behavior as an important argument. We inserted that insinuation expecting it to exert a damaging effect on the plaintiff’s case, but virtually no jurors even mentioned it as a factor vote. Indeed only one juror finding for the defendant mentioned the plaintiff’s history of drug use, while three jurors finding for the defendant affirmatively denied it had anything to do with drug seeking behavior.

The CRM data, however, tell a very different story. The Bollinger band analytic approach flagged portions of the transcript discussing drug use as exerting a substantial impact (see Figure 9), and additional exploration in terms of cultural cognition suggested the effect is driven by people with a hierarchist-communitarian disposition (see Figure 10). These insights would have been altogether missed had we relied solely on non-CRM data.

Our examples have illustrated two of the three empirical designs to which CRM can be usefully coupled (see Table 1). The first—and most obvious—is to use CRM in a within-subjects design. Here we simply assess changes in responses over time, correlated with changes in the content of the trial. Causal inference is most perilous in this context, however, since we lack a counterfactual for comparison. One cannot say
that a given part of the transcript caused an observed change in response unless one assumes that no such change would have been observed in the absence of that segment. But what is the imagined counterfactual? A blank screen? The overall average score? These sorts of questions counsel caution in using CRM in its most obvious application.

The caution we recommend is to implement a statistical decision rule, such as the 60MA ± (60SD×2) bands we deployed. This does not obviate the need to think about the data and use subjective judgment, but it does help provide transparency in how one has approached the data. If a segment within the bands is picked out for discussion, or a segment outside the bands ignored, then the researcher is at least prompted to explicitly defend that decision (i.e., to explain why an exception to the statistical decision rule is sensible), thereby empowering the reader to better critique the analyses.

Table 1. Potential experimental designs using CRM

<table>
<thead>
<tr>
<th>Design</th>
<th>Allows</th>
<th>Worry</th>
</tr>
</thead>
<tbody>
<tr>
<td>within subjects, correlating changes over time to stimulus segments</td>
<td>hypothesis generation</td>
<td>lack of counterfactual; changes may be relative to prior positions; varying variance</td>
</tr>
<tr>
<td>between subjects, split on IVs or other DVs</td>
<td>weak causal inference</td>
<td>confounded by unknown third variable</td>
</tr>
<tr>
<td>between subjects, randomized, controlled trial (RCT)</td>
<td>strong causal inference</td>
<td>gold standard for rigor, but can be cumbersome to design and field iterative RCTs for each hypothesis</td>
</tr>
</tbody>
</table>
The second use is an observational between-subjects approach. Here we split the CRM data on some other observable, such as final verdict or cultural cognition score. When we observe points of inflection—where one group of subjects is reacting differently than another group of subjects—we can infer that it was the proximate portion of the stimulus (interacting with the splitting variable) that caused the change. In this instance one group provides the counterfactual point of comparison for the other, thereby empowering a modest causal inference that the splitting variable is having some impact on the differential responses. The inference is only “modest” because of the usual cautions associated with comparing groups split on observed characteristics: without the benefit of random assignment, we worry that unobserved variables might be driving the observed results. We assumed cultural cognition could explain the differential responses to insinuated drug-seeking behavior, but maybe the real driver is a third variable itself correlated with cultural cognition—age, religious belief, or something else of which we are unaware.

The final method is coupling CRM with a between-subject, randomized experimental design. One could, for example, experimentally manipulate the videotaped trial stimulus into treatment and control conditions, as we did here. This use of CRM allows one to not only see the immediate impact of a given piece of testimony but also see how it changes the subjects’ receptiveness to subsequent testimony, given a comparison to a control condition where the segment was omitted. This is precisely the method we use in Yokum and Robertson (2014).
The field is open and the time ripe for further use of CRM in jury research. For example, to our knowledge, there are no prior studies coupling CRM with experimental designs. Our hope is that the discussion here, and some of the analytic tools we provide, will spur research to fill this gap.
### Attachment D. Transcript excerpts

<table>
<thead>
<tr>
<th>Min Seg</th>
<th>To</th>
<th>Size</th>
<th>Relevant Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 DO</td>
<td>DEF</td>
<td>4.40</td>
<td>The evidence will show that Mr. Stevens’s complaints could have been due to muscular soreness, which usually goes away on its own. The medical records show that Dr. Dennis told Mr. Stevens to come back if the pain got worse, but Mr. Stevens didn’t do that for several months. Finally, our expert witness, Dr. Davidson, will testify that Dr. Dennis met the standard of care, and did exactly what any good doctor would do in a case like this. Thank you for your time.</td>
</tr>
<tr>
<td>8 PEB</td>
<td>PL</td>
<td>3.49</td>
<td>The blinds help minimize the ability of the manufacturer to bias the outcomes of the study, and in a similar way in litigation some experts work under blinds to make sure that the litigants are unable to hand-pick favorable experts so that the experts are more neutral. Q. So on the one hand it prevents the litigant from hand-picking an expert. Q. Oh, I see. Did it also affect the way you rendered your opinion? A. Yes, I did not know who was asking for the opinion. So you performed as a blind expert in this case? A. I did.</td>
</tr>
<tr>
<td>14 PES</td>
<td>PL</td>
<td>3.84</td>
<td>Q. And what would you have concluded from these two findings? A. Well, both of those findings are very significant because they both show-, they’re very suggestive of the nerve being pinched at the level where it comes out of the backbone. Q. And what did Dr. Dennis find? A. He thought this was consistent with non-specific low back pain and he prescribed some anti-inflammatory medications and told the patient to follow up as needed. Q. And can you tell us what happened next? A. Well, I think the patient went for some physical therapy, did that for three months but the pain was getting worse and when he-, after three months re-presented, he had actually worse symptoms.</td>
</tr>
<tr>
<td>16 PES</td>
<td>PL</td>
<td>2.77</td>
<td>A. The MRI showed that there was disc herniation, which is when the disc is pushed out and there’s actually a large fragment pushing on the nerve that goes down from the backbone into the lower leg, which was pushing on the nerve. Q. So this finding was, just to be clear, three months after Dr. Dennis had actually examined Mr. Stevens? A. Correct.</td>
</tr>
<tr>
<td>17 PES</td>
<td>DEF</td>
<td>3.55</td>
<td>I think there are some figures that show that eighty percent of adults will have low back pain at some point in their life. But there are a number of symptoms that are suggestive of more significant low back pain that needs to be acted on.</td>
</tr>
<tr>
<td>18 PES</td>
<td>PL</td>
<td>5.16</td>
<td>And so if Dr. Dennis had done that when the patient presented with neurologic signs, the patient would have been saved three months of, of delay. Q. So, so you’re saying that he may not be disabled today if the standard of care had been met? A. Correct.</td>
</tr>
<tr>
<td>20 PEX</td>
<td>DEF</td>
<td>9.15</td>
<td>Q. So doctor, just to summarize what we see in the medical records. Dr. Dennis did tell Mr. Stevens to come back if the pain got worse, isn’t that right? A. The medical records do reflect that. Q. And the pain did get worse, but Mr. Stevens didn’t come back for three months, isn’t that true? A. That is true. It’s not clear to me how forcefully Dr. Dennis made it clear...</td>
</tr>
</tbody>
</table>
5. Using CRM in jury research

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>how important that was, but it does say in the medical records that he told him to call him and he didn’t.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>DEF</td>
<td>8.20</td>
</tr>
<tr>
<td>Q: Isn’t it true that he had low back pain and right shoulder pain from a past motor vehicle accident?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>PL</td>
<td>5.45</td>
</tr>
<tr>
<td>However, given that there were signs on the exam of neurologic compromise which was different from previous exams in this case, I think that that would trump any concern that one would have about drug-seeking behavior.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>PL</td>
<td>4.01</td>
</tr>
<tr>
<td>No, I think that the, the context, the exam, the history were all concerning enough to be aggressive in this case.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>DEF</td>
<td>3.54</td>
</tr>
<tr>
<td>Q. What do you mean by conservative management? Tell me a little more.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Well, when a patient comes to your office complaining of something like low back pain, the initial management can either be to, you know, immediately go to doing an MRI that could then lead to surgery or to try more conservative measures first to see if they relieve the symptoms such as physical therapy and anti-inflammatory drugs and otherwise, you know, getting rid-, sort of not working as hard as you might have before.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>PL</td>
<td>2.79</td>
</tr>
<tr>
<td>A. Well, yes, those are signs of Cauda equina syndrome, but I mean, I think that they’re, you know, you sort of, when you do a neurological exam, those are parts of the neurological exam that you do and they were normal. Q. But you can’t use those facts to rule out radiculopathy, the actual injury that Mr. Stevens has here, right?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>PL</td>
<td>3.46</td>
</tr>
<tr>
<td>Q. I also want to ask about, I think in your testimony you relied on the fact that there wasn’t a single, I think you called it inciting event that caused this pain.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>PL</td>
<td>2.85</td>
</tr>
<tr>
<td>Q. So, are you familiar with the terms “selection bias” and “affiliation bias” and “compensation bias”? A. Yes, I’ve heard of those terms. Q. And isn’t it true that the purpose of using a blind expert like the plaintiff’s expert witness is to avoid those potential biases, isn’t it? A. That’s what he described but I don’t really have anything to do with that.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>DEF</td>
<td>3.26</td>
</tr>
<tr>
<td>As I said before, in deciding the facts of this case, you should consider what testimony to accept, and what to reject. I will now instruct you as to the law. First, the Plaintiff, Mr. Stevens bears the burden of proof.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


Author Note

This work was done in collaboration with Christopher T. Robertson.

This research was supported in part by grants from the Petrie Flom Center at Harvard Law School and the Faculty Allotment at the University of Arizona. Thanks to Gregory Schwartz, Aaron Kesselheim, and Tom Mauet for serving as actors and consultants, and to Germar Townsend, Tess Gemberling, Carol Ward, Judy Parker, Barbara Lopez, and Bert Sky for research and administrative support. Thanks also to Guy Hagen and Watchfire Analytics for providing access to continuous response measurement technology.
Abstract

We have earlier proposed, and tested with an online sample, an institutional solution—blinded expertise—that seeks to provide jurors with expert witnesses who are less biased and, therefore, rightly more persuasive. The central idea is to have an intermediary select an expert witness and then have that expert render an opinion before learning which litigant has requested the opinion. Here we replicate the finding with an in-person community sample ($N = 95$), showing that the plaintiff is 20-percentage points more likely to win, 95% CI [.01, .39], $p = .04$, if his witness is a blinded rather than hand-picked expert. We also couple the experimental design with the use of continuous response measurement (CRM), in order to probe the mechanisms behind jurors perceptions of expert witness credibility and, in particular, how blinding affects those perceptions. We find, among other things, that blinding affords both a direct benefit to the litigant—the mere fact a blinded witness is used seems to win points—as well as downstream benefits, in that later direct examination of the blinded witness is weighted more favorably and, likewise, he is immunized from some of the damage otherwise done during cross examination.

Key words: jury research, credibility, scientific reasoning, continuous response measurement (CRM), medical malpractice
Background

Jurors confronted with expert witness testimony face a difficult challenge. Professor Gross has referred to it as the “essential paradox” of expert testimony. Namely, “[w]e call expert witnesses to testify about matters that are beyond the ordinary understanding of lay people (that is both the major practical justification and a formal legal requirement for expert testimony), and then we ask lay judges and jurors to judge their testimony” (Gross, 1991, p. 1182). Compounding the challenge is that, in the American system, “expert witnesses are hand-picked by the litigants, coached toward favorable opinions, and compensated for continuing work only as long as those opinions are favorable” (Robertson & Yokum, 2012, p. 766; see also Robertson, 2010). The end result is that jurors often see expert witnesses from the respective litigants who adamantly insist on the truth of two positions that are diametrically opposed.

We have proposed and tested an institutional solution—blinded expertise— that seeks to provide jurors with expert witnesses who are less biased and, therefore, rightly more persuasive (Robertson, 2010; Robertson & Yokum, 2012). The central idea is to have an intermediary select an expert witness and then have that expert render an opinion before learning which litigant has requested the opinion. In this fashion, the expert is blind as to which side he or she is serving and, therefore, also shielded from exposure to many of the financial and social pressures to provide an opinion favorable to the requestor (see Robertson, 2010, for an extensive discussion). This institutional fix to avoid bias is only part of the solution, however; in order for its benefits to be felt in the courtroom, jurors must actually find blinded experts more persuasive than
handpicked and coached experts. If jurors are not persuaded, then all the extra effort to blind is for naught.

Robertson and Yokum (2012) report an initial test of the impact of blinded experts on juror verdicts. We cast 275 jury-eligible participants as jurors in a medical malpractice case. Using a 3×2 between subjects design, we carefully manipulated the trial videotape such that neither expert was blinded, only the plaintiff’s expert was blinded, or only the defendant’s expert was blinded; this was crossed with whether or not a special jury instruction pertaining to blinding was administered. The effect of blinding an expert was substantial. “Blind experts, testifying on either side, were perceived as significantly more credible, and were more highly persuasive, in that they doubled (or halved) the odds of a favorable verdict, and increased (or decreased) simulated damages awards by over $100,000” (Robertson & Yokum, 2012, p. 765).

Our aim here is to replicate and extend this finding. Replication is important. Scientists have always duly nodded their head to this statement, but only recently has the true weight of the responsibility to replicate come to be felt. A growing realization that much published research fails to replicable (see e.g., Yong, 2012), coupled with a growing awareness that null hypothesis testing and its resultant \( p \) value can be deeply misleading (e.g., S. N. Goodman, 1992; Cumming, 2008; Nuzzo, 2014), are some of the primary driving forces behind the renewed focus on replication. Henry Roediger, past President of the American Psychological Society, puts the upshot well:

To those who argue that a robust level of statistical significance is all one needs to assure replicability, I recall the aphorism (attributed to Confucius) that “One replication is worth a thousand t-tests.” Words to live by. And if we replicate
our results routinely, we do not need to worry so much about the poor logic of
null hypothesis statistics or using Bayesian statistics to try to determine what
happened in a single experiment or study. If you obtain an effect, just replicate
it (perhaps under somewhat different conditions) to be sure it is real (Henry L.

We aim to do precisely as Prof. Roediger advises.

We also extend the prior study in three important ways. First, we use a very
different sample, this time a community sample from the metropolitan area of Tucson,
AZ, as opposed to an online sample drawn from Amazon Mechanical Turk (“mTurk”).
Many basic psychological findings have been replicated using online samples such as
mTurk, but there are important exceptions. Researchers are vigorously debating the
pros and cons of mTurk and its ilk, trying to delineate when it is or is not an appropriate
to use (see e.g., Mason & Suri, 2012; Crump, McDonnell, & Gureckis, 2013; J. K.
Goodman, Cryder, & Cheema, 2013; Kahan, 2013). We contribute to this debate by
testing the same trial stimuli in a wholly different population and setting.

Second, and perhaps more notably, we use continuous response measurement
(CRM) to probe in more detail how blinding an expert impacts juror reasoning. CRM
entails the continuous measurement of verdict preference over time. Consider, for
example, the following issue. Are jurors directly impressed that a litigant has brought a
blinded expert, in the sense that the testimony about blinding directly tilts verdict
preferences in that party’s favor? Or alternatively does blinding color the way jurors see
subsequently presented evidence?

Third is we had jurors proceed to deliberate in groups. Deliberations were
recorded, in order to probe how the blind expert concept played out during
deliberations. This provides a window into how jury (as opposed to juror) verdicts are affected, although we are too under-powered to say anything with confidence on this particular point. This qualitative analyses of results will be reported elsewhere.

**Method**

**Participants**

Ninety-five jury eligible adults were recruited from the Tucson, AZ metropolitan area. A fifteen dollar gift card was provided as compensation. Participants were told generally that the study concerned a mock trial, but were not given specifics as to the hypotheses being tested or their assignment into experimental conditions. As can be seen in Table 1, the overall sample roughly mirrors the population characterized by the U.S. Census in terms of education, gender, age, and race (minus a tilt toward being younger and more educated), and the Control and Plaintiff BE sub-samples are demographically indistinguishable. None of these demographic variables predicted verdict outcomes, and we did not expect (nor were there) any interactions with the independent variables under study.

**Trial stimulus and procedure**

Participants were cast as jurors in a medical malpractice case, using a 35-45 minute mock trial video as stimulus. The video was edited, in 2×1 between-subjects design, such that either: (a) both expert witnesses were hand selected by their respective counsel (Control); or (b) the plaintiff’s expert was a blinded expert while the defendant’s expert remained hand selected (Plaintiff BE). The primary outcome was
verdict preference, and a variety of secondary measures associated with perceptions of expert witness credibility were also collected.

Table 1. Demographics

<table>
<thead>
<tr>
<th>Education</th>
<th>Neither BE (n = 47)</th>
<th>Plaintiff BE (n = 48)</th>
<th>Total (N = 95)</th>
<th>U.S. Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; HS Diploma/GED</td>
<td>4%</td>
<td>0%</td>
<td>2%</td>
<td>18%</td>
</tr>
<tr>
<td>HS Diploma/GED</td>
<td>4%</td>
<td>17%</td>
<td>11%</td>
<td>30%</td>
</tr>
<tr>
<td>Some College/Assoc.</td>
<td>40%</td>
<td>40%</td>
<td>42%</td>
<td>27%</td>
</tr>
<tr>
<td>College Grad</td>
<td>26%</td>
<td>19%</td>
<td>23%</td>
<td>17%</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>19%</td>
<td>21%</td>
<td>21%</td>
<td>10%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38%</td>
<td>44%</td>
<td>41%</td>
<td>49%</td>
</tr>
<tr>
<td>Female</td>
<td>55%</td>
<td>52%</td>
<td>54%</td>
<td>51%</td>
</tr>
<tr>
<td>Age Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>21%</td>
<td>29%</td>
<td>25%</td>
<td>13%</td>
</tr>
<tr>
<td>25-34</td>
<td>23%</td>
<td>17%</td>
<td>16%</td>
<td>18%</td>
</tr>
<tr>
<td>35-44</td>
<td>15%</td>
<td>4%</td>
<td>8%</td>
<td>19%</td>
</tr>
<tr>
<td>45-59</td>
<td>15%</td>
<td>23%</td>
<td>19%</td>
<td>27%</td>
</tr>
<tr>
<td>60+</td>
<td>19%</td>
<td>23%</td>
<td>21%</td>
<td>23%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>77%</td>
<td>73%</td>
<td>75%</td>
<td>74%</td>
</tr>
<tr>
<td>Non-White</td>
<td>33%</td>
<td>27%</td>
<td>25%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Note: Five participants failed to provide age, gender, and education data; percentages thus do not add to 100%.

Full details about the trial stimulus are provided in Robertson and Yokum (2012), where the same case was presented to an online sample of 275 mock jurors recruited through Amazon Mechanical Turk. Briefly though, we used two of the five conditions: the control and Plaintiff BE condition. Participants in the Control condition viewed a 35-
minute video containing instructions from the judge (modeled after the Revised Arizona Jury Instructions), opening statements from attorneys for the plaintiff and defendant, examination and cross-examination testimony of the two expert witnesses (one for the plaintiff and one for the defendant), closing statements, and then further instructions from the judge. The survey was hosted on surveymonkey.com, although participants actually completed the task in a computer laboratory at the law school, sitting at individual computer stations.

The trial itself concerned the failure of a primary care physician to diagnose a possible case of lumbar radiculopathy and refer the patient to imaging, which allegedly would have allowed timely surgery and avoidance of the permanent disability that the patient now suffers. The primary dispute concerned whether the physician-defendant met the standard of care when, instead of ordering imaging, he simply instructed the patient to take pain-killers and return if the pain got worse.

The Plaintiff BE condition was created by splicing into the video eight additional video segments, amounting to 10 minutes of footage discussing the blind expertise concept. For example, several minutes were added to the direct examination of the plaintiff’s expert wherein the attorney-expert colloquy described how that expert was randomly selected by a third-party intermediary, and that the expert rendered his opinion before knowing which party requested the consultation. In this manner the expert witness for the plaintiff was transformed into a blind expert, while the substantive testimony, statements, and instructions remained exactly identical as in the Control condition. Given the prior finding that the Defendant BE condition included in
Robertson and Yokum (2012) mirrored the effect of the Plaintiff BE condition (and the jury instruction manipulation was ineffectual), we dropped the former in order to increase the sample sizes available for the Control and Plaintiff BE conditions.

**Continuous response measurement (CRM)**

Unlike our prior study (Robertson & Yokum, 2012), we also used here a technique of continuous response measurement (CRM). The basic idea is straightforward: participants, rather than only providing a verdict judgment at trial conclusion, also indicate their moment-to-moment verdict preferences continuously throughout the duration of the trial. There are multiple ways to collect and analyze such data. We develop and explain CRM at length in Yokum and Robertson (in prep).

In this case, participants were provided a sliding scale positioned onscreen just below the stimulus video (see Figure 1), which captured ratings from zero to 100 in half-second intervals. They were instructed (and taught via tutorial from a research assistant) to: “Move the ball to indicate the degree to which you think the defendant or the plaintiff should win the case, based on what you have heard so far. Evaluate constantly as you watch the trial.” The indicator would remain in place between juror ratings (rather than “snapping back” to the center, an alternative implementation). Thus the recordings represent cumulative CRM, meaning that the rating at any given moment expresses the overall verdict preference of the juror. It is, to put it differently, how the participant would score the case if the trial were to abruptly stop, and thus it can be used to predict the participant’s vote at any given point in time (see Yokum &
Robertson, in prep, for detailed discussion, including the alternative option of momentary CRM).

Figure 1. User interface for collecting CRM data.

**Verdict and monetary damages questionnaire**

After viewing the trial stimulus, all participants were asked to render a binary verdict, answering “Yes” or “No” to the question, “Based on the instructions provided by the judge in the video, do you believe that the Plaintiff has proved, by the greater weight of the evidence, that the Defendant committed medical negligence?” Subsequent questions asked for a verdict rating on a 6-point Likert scale, from 1 (clearly not medical negligence) to 6 (clearly medical negligence), and—if finding for the plaintiff—an assignment of monetary damages for pain and suffering.

**Credibility questionnaire**

Participants then indicated, on a 4-point Likert scale from 1 = strongly disagree to 4 = strongly agree), whether each expert witness was “knowledgeable,” “logical,” “clear,” “trustworthy”, “fair,” and “honest.” Overall credibility is simply the sum of all six trait ratings. But research also suggests that credibility is usefully thought of as a construct composed of multiple factors (2012, pp. 783–785). The factors mapped here are skill (the sum of knowledgeable, logical, and clear scores) and genuineness (the sum
of honesty, trustworthy, and fair scores). Skill captures the expert’s technical training or intellect, while genuineness captures a sense of believability. As we put it before:

To appreciate the possible difference between these two factors, you might image, for example, an expert with outstanding credentials (but who seems to be lying). This expert would have a high skill score (he [] could give an accurate opinion), but a low genuineness score (since he lies). On the other hand, an expert might appear to be doing his honest best to give an accurate opinion, but nonetheless seem unqualified in training or intellect. This expert would have a high genuineness score, but a low skill score. The point is that both sorts of considerations (and possibly others) factor into an overall assessment of credibility, and different factors might be differentially sensitive to the blind expert concept (Robertson & Yokum, 2012, p. 785, n. 17).

Finally they completed demographic, Need for Cognition (NFC), and Cultural Cognition (CC) questionnaires.

**Self-reports and deliberation**

After rending their verdicts and (if applicable) monetary damages awards, participants were also asked to explain the reasoning behind their decisions, via a short written prompt. (“In your own words, please provide a sentence or two explaining your decision on this page.”) After completing their questionnaires, they were also assembled into groups for a recorded deliberation stage. There were 18 jury groups in total, each allowed to deliberate for about 45 minutes.

**Results**

**Verdict**

Referring to Table 2, 30% of participants found in favor of the plaintiff when both expert witnesses were hand selected. When the plaintiff used a blinded expert, on the
other hand, the plaintiff’s win rate increased substantially, to 50%—an absolute increase of 20%, 95% CI [.01, .39], \( p = .04 \).

Table 2. Verdicts and verdict rating. Frequency (and percentage) of verdicts in favor of the plaintiff, along with mean certainty scores on a 6-point Likert scale (and standard deviation). Compared to the control condition where 30% of respondents voted for the plaintiff, the plaintiff’s unilateral use of a blind expert increased the plaintiff win rate to 50%—a 20% absolute increase, 95% CI [.01, 39], \( p = .04 \).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Verdict for Plaintiff (%)</th>
<th>Mean Rating (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE (control)</td>
<td>14 (30%)</td>
<td>3.13 (1.47)</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>24 (50%)</td>
<td>3.42 (1.50)</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>38 (40%)</strong></td>
<td><strong>3.27 (1.48)</strong></td>
</tr>
</tbody>
</table>

Table 3 provides a more precise estimate using logistic regression, including the introduction of demographic covariates as further control (beyond randomization). Coefficients are translated into both original probability units (with 95% CI) and odds ratios (with 95% CI) to aid interpretation of effect sizes and precision. Results are coded such that a positive \( \beta \) indicates a verdict in favor of the plaintiff is more likely, with Plaintiff BE (0 = control; 1 = plaintiff uses blinded expert), Age10 (age in decade intervals), male (0 = female; 1 = male); white (0 = not white; 1 = white), and college (0 = less than college; 1 = college degree of higher).
Table 3. Logistic regression predicting verdict in favor of plaintiff.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$B$</th>
<th>S.E.</th>
<th>$p$</th>
<th>95% CI for Probability Increase</th>
<th>95% CI for Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
<td>%</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.36</td>
<td>0.73</td>
<td>.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>0.99</td>
<td>0.46</td>
<td>.03</td>
<td>.02</td>
<td>.24</td>
</tr>
<tr>
<td>Age10</td>
<td>-0.02</td>
<td>0.14</td>
<td>.87</td>
<td>-.07</td>
<td>-.01</td>
</tr>
<tr>
<td>Male</td>
<td>-0.10</td>
<td>0.48</td>
<td>.83</td>
<td>-.26</td>
<td>-.03</td>
</tr>
<tr>
<td>White</td>
<td>-0.86</td>
<td>0.60</td>
<td>.15</td>
<td>-.50</td>
<td>-.22</td>
</tr>
<tr>
<td>College</td>
<td>0.61</td>
<td>0.57</td>
<td>.29</td>
<td>-.12</td>
<td>.15</td>
</tr>
</tbody>
</table>

Again the substantial impact of a blinded expert is observed, and this more precise estimate indicates an even larger effect: a 24% absolute increase in probability, 95% CI [.02, .47], $p = .03$, or equivalently a 2.70 odds ratio, 95% CI [1.11, 6.82], which is to say the plaintiff is about 170% more likely to win if using a blinded rather than handpicked expert witness. Notably, Robertson and Yokum (2012, p. 776) observed a similar odds ratio of 2.18, 95% CI [1.01, 4.70].

Case expected value

Table 4 provides descriptive statistics on case expected value (CEV). CEV is a function of both the underlying verdict rate and positive damage awards (specifically, defense verdicts are included as zeros in all computations). The data are, as is common for damages awards, highly right skewed, with a handful of severe outliers, and constituted by a large number of zeros. This renders the mean potentially misleading, and familiar parametric statistical approaches (that is, methods that assume a normal,
or roughly bell-shaped, distribution), such as a t-test, inappropriate. The 10% trimmed mean, calculated by removing the smallest 5% and largest 5% of scores, is a more robust point estimate of central tendency (Wilcox, 2010, pp. 131–136). By this simple measure the blinded expert exerts a powerful effect: the 10% trimmed more than doubles, from $63,154 (Neither BE) to $151,227 (Plaintiff BE)—a difference of about $88,000, or a relative 239% increase. A Wilcoxon rank sum test, a non-parametric statistical test that makes no assumptions about the underlying distribution shape, confirms that this is a reliable difference, \( W = 698, p = .02 \). (Note this test uses all the data, not the trimmed scores; the latter is simply to describe the central tendency).

**Table 4. Case Expected Value.** Pain and suffering awards (U.S. dollars), including defense verdicts as zeros.

<table>
<thead>
<tr>
<th>Condition</th>
<th>( n )</th>
<th>Median</th>
<th>Mean</th>
<th>10% Trim ( M )</th>
<th>( SD )</th>
<th>Wilcoxon rank sum test on difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE (control)</td>
<td>47</td>
<td>0</td>
<td>131,698</td>
<td>63,154</td>
<td>384,431</td>
<td>( W = 698, p = .02 )</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>48</td>
<td>10,000</td>
<td>248,896</td>
<td>151,227</td>
<td>579,118</td>
<td></td>
</tr>
</tbody>
</table>

**Award hedging**

A different question is whether jurors finding for the plaintiff “hedge” against uncertainty by awarding a lower damages award (Robertson & Yokum, 2012, pp. 778–779). The proposition is that participants are forced to make a binary choice, and beyond that, the only way to express uncertainty is to discount the damages award. To put it more formally, jurors finding the preponderance of evidence indicates medical malpractice (in theory \( p_{\text{plaintiff}} > 0.5 \)) render a verdict for the plaintiff; their precise \( p_{\text{plaintiff}} \)
estimate then affects the damages awards, such that $p_{\text{plaintiff}} = 1.0$ results in $X$ while $0.5 < p_{\text{plaintiff}} < 1.0$ results in $X$ discounted by some factor $g$. (The properties of $g$ are open to empirical investigation; it is probably not a straightforward linear function, if it exists).

This inference is possible here because neither expert witness discussed pain and suffering. Their testimonies focused exclusively on whether the standard of care had been meet (and thus whether the defendant should be liable), without any commentary on the severity of the plaintiff’s pain and suffering.

Table 5 shows descriptive statistics of damages awards, conditional on a finding in favor of the plaintiff (that is, excluding defense verdicts). Jurors finding for the plaintiff in the Plaintiff BE condition awarded a smaller amount ($M_{10\% \text{ Trim}} = $424,632) than jurors finding for the plaintiff in the Neither BE condition ($M_{10\% \text{ Trim}} = $457,750), although this difference is not statistically reliable (Wilcoxon rank sum test: $W = 121$, $p = .87$). By this metric there no support for a “hedging” effect. On the other hand, the sample here is so small that the test is underpowered, so we cannot confidently rule out the possibility of an effect either.
Table 5. Positive Damages Awards. Pain and suffering awards (U.S. dollars), conditional on a verdict in favor of the plaintiff.

<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>Median</th>
<th>Mean</th>
<th>10% Trim M</th>
<th>SD</th>
<th>Wilcoxon rank sum test on difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither BE (control)</td>
<td>47</td>
<td>87,500</td>
<td>457,750</td>
<td>457,750</td>
<td>635,903</td>
<td>(W = 121, p = .87)</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>48</td>
<td>100,000</td>
<td>527,062</td>
<td>424,632</td>
<td>765,042</td>
<td></td>
</tr>
</tbody>
</table>

**Expert credibility assessments**

Table 6 catalogues overall assessments of expert witness credibility, as well as deconstructed in terms of “skill” and “genuineness” (see Methods above, p. 246). It replicates the pattern observed by Robertson and Yokum (2012, p. 784). The plaintiff’s expert, if blinded, enjoys a substantial boost in overall credibility (from 6.17 to 6.93, \(t(91) = 3.76, p < .01\)), reflecting meaningful gains in both perceived skill (3.14 to 3.46, \(t(92) = 2.97, p < .01\)) and perceived genuineness (3.03 to 3.47, \(t(91) = 4.18, p < .01\)). As for the defendant’s expert, his overall credibility dips from 5.84 to 5.65, although our confidence in a reliable difference is low, \(t(88) = -0.80, p = .42\). This reflects a drop in perceived genuineness (from 2.82 to 2.62, \(t(90) = -1.63, p = .11\)), while perceived skill remains constant around three points (from 3.01 to 3.03, \(t(92) = 0.17, p = .87\)).
Table 6. Expert Credibility Assessments. Mean (standard deviation) assessments of skill and genuineness, on 4-point Likert scales from 1 (strongly disagree) to 4 (strongly agree) when neither (Neither BE) or only the plaintiff (Plaintiff BE) retains a blind expert. Overall credibility is the sum of skill and genuineness. Transforming the plaintiff’s expert into a blinded expert increased his credibility (in both skill and genuineness terms), while also potentially decreasing the perceived genuineness of the other expert.

<table>
<thead>
<tr>
<th>Blind Expert</th>
<th>Plaintiff’s Expert</th>
<th>Defendant’s Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Credibility</td>
<td>Skill</td>
</tr>
<tr>
<td>Neither BE</td>
<td>6.17 (1.04)</td>
<td>3.14 (0.54)</td>
</tr>
<tr>
<td>Plaintiff BE</td>
<td>6.93** (0.93)</td>
<td>3.46** (0.50)</td>
</tr>
</tbody>
</table>

* Planned contrast relative to Neither BE condition is significant at $p < .05$ (or ° if $p < .10$).

Continuous Response Measurement (CRM)

Although the concept of CRM is straightforward, analyzing the results in a robust and objective fashion is not. Figure 2 shows the mean rating over time split by both experimental condition. A qualitative eyeballing of the chart reveals some notable trends, four of which are highlighted in boxes A, B, C, and D. Consider each in turn.

First, jurors render similar ratings up until the 7th minute, at which point the plaintiff enjoys a slow upward march in support. This boxed area (A) constitutes the plaintiff’s direct examination of his own expert witness, and in particular the dialogue wherein details about the blinding process are being explained. The blue control line is perfectly flat because, recall from the methodology, there is no video footage at this point for controls—the line is merely extrapolating over the blank space using the adjacent input points.
6. CRM analysis of blinded experts

Between the section demarcated by box A and box B, ratings seem to roughly rise and fall together.

And then something dramatic happens in section B. Note that this portion of the trial is identical for treatment and control subjects. Participants who were introduced to the plaintiff’s expert witness as a blinded expert apparently respond differently to whatever is happening here. The plaintiff’s expert is being crossed by the defense attorney at this moment, in particular about the fact that the defendant Dr. Dennis
instructed the plaintiff Mr. Stevens to return if the pain worsened, but he failed to do so:

[Defendant’s attorney]: . . . Dr. Dennis did tell Mr. Stevens to come back if the pain got worse, isn’t that right?

[Plaintiff’s expert witness]: The medical records do reflect that.

[Defendant’s attorney]: And the pain did get worse, but Mr. Stevens didn’t come back for three months, isn’t that true?

[Plaintiff’s expert witness]: That is true.

Box C illustrates another segment of identical video across conditions, but this time with the treatment and control groups seemingly rising and falling together. The lines are not exactly parallel, and at certain points diverge in opposite directions; but it is difficult to know if any such shifts are reliably different, as opposed to random noise, just from looking.

Box D captures ratings when the trial ends. Examining whether jurors’ final ratings were above or below the 50% line, we find that 50% of subjects in the treatment condition and 30% of subjects in the control condition were above the midline. Notably, this matches to the digit the rates observed on the subsequent binary verdict questionnaire.

Figure 3 charts mean CRM split by both condition and final verdict, with a box superimposed around the 21st minute—it is labelled Box B again, to maintain consistency with the labelling from Figure 2. The top dark green line at the end of the trial, for example, shows ratings over time for those jurors who were in the control condition and ultimately found that the defendant was liable for medical negligence.
Note how up until the 20th minute, jurors in the Plaintiff BE condition, whether ultimately finding for the plaintiff or defendant, follow a similar trend. Intriguingly, the dialogue in Box B is splitting jurors in the treatment condition: for those who will find for the plaintiff there is no impact, while for those who will ultimately side with the defendant, it seems this is the decisive moment. It is, as just discussed, the moment when patient responsibility is being discussed, that is, the failure of the patient to return as instructed when his pain worsened.

The persuasive power of the blinded expert is evident. In those segments unique to the treatment condition (where the green lines are flat due to no video, e.g. between minutes 7-11 or 35-42), we can clearly see verdict preferences tilting further and further in favor of the plaintiff. Jurors are impressed with experts selected via a blinded mechanism. Note that, during these moments, it is simply the mechanism of blinding that is being discussed—not, in contrast, anything about the substance of the case. It seems jurors are crediting points, so to speak, to the plaintiff merely for the fact of bringing a blinded expert to the stand.
To more formally analyze these results, we apply a method developed in Yokum and Robertson (in prep). In particular, we compute a 60-second moving average (60MA) as well as a 60-second moving standard deviation (60SD), and designate as statistically significant any mean ratings that breaks higher or lower than 60MA ± (60SD×2). This tests whether the mean score at any given moment would be unlikely given the average mean and variance observed over the prior sixty seconds. Figure 4 and Figure 5 chart the mean CRM in the control and treatment conditions, respectively, coupled with 60MA ± (60SD×2) bands.
6. CRM analysis of blinded experts

Figure 4. CRM chart of mean rating (blue line), plus 60MA (dark orange line) and ± (60SD×2) bands (yellow lines) in the control condition.

Figure 5. CRM chart of mean rating (blue line), plus 60MA (dark orange line) and ± (60SD×2) bands (yellow lines) in the treatment condition.
Most of the trial segments flagged by our analytic method are identical across conditions. Figure 6 identifies, on the treatment CRM chart, events that are flagged as statistically different across both conditions. (We could have just as easily shown this on the control chart, but had to pick one or the other). Both opening statements had a substantial impact, for example, as did testimony regarding topics such as: the plaintiff’s expert’s testimony regarding the plaintiff’s pain or that, in his professional opinion, the standard of care had been violated; the plaintiff never returned to see the doctor, despite instruction to do so if the condition worsened; and so forth.

Figure 7, on the other hand, highlights events that were flagged in the control condition but not flagged in the treatment condition. An interpretive challenge here is assessing whether these non-significant results reflect jurors who truly perceive the evidence differently, having been exposed to a blinded expert, versus a null that is due to high variance that was perpetuated by an earlier high impact event. In particular, one might worry about a sort of masking effect, wherein a large effect at $t_1$ temporarily inflates the 60SD, and as a result the same sized effect at $t_2$ no longer exceeds 60SD×2. For example, in the control condition, the mean rating significantly jumps from about 49 to 51 when the plaintiff’s expert mentions a neurological exam positively indicated a problem; in the treatment condition, the jump is actually the same (from 55 to 57), but it (unlike in the control condition) comes off the heels of the blinded expert testimony—the effect of which was large and therefore temporarily inflated the downstream 60SD. Of course, as only a 2-point shift, neither is likely substantively significant (even if statistically significant), but the point is the possibility of masking.
Figure 6. Events flagged as statistically significant in both the control and treatment groups.

Figure 7. Events that were flagged as statistically significant in the control group, but not in the treatment group.
Discussion

The blinded expert is, once again, found to be an extremely persuasive individual within the courtroom. The controlled experimental design used here meant that all jurors heard exactly the same substantive testimony. The only difference was whether, via careful video editing, we re-cast the plaintiff’s expert witness as a blinded expert. As a blinded expert, his words carried substantially more weight. The plaintiff win rate increased from 30% to 50% when his expert witness was employed via a blinded mechanism rather than the tradition handpick-and-coach method. This represents a 24% absolute increase in probability, 95% CI [.02, .47], or equivalently a 2.70 odds ratio, 95% CI [1.11, 6.82]. It also constitutes a replication in line with the large effect size observed in our earlier online experiment, which was an odds ratio of 2.18, 95% CI [1.01, 4.70] (Robertson & Yokum, 2012, p. 776).

The replication here is particularly notable on account of the substantial difference in underlying verdict win rates. Robertson and Yokum (2012), using an online mTurk sample, found that the plaintiff won 46% of the time when neither expert was blinded (control) and 62% of the time when only the plaintiff’s expert was blinded (treatment). We observe here much lower rates across the board: 30% in the control condition, and 50% in the treatment condition. Such large base rate swings suggest that non-evidentiary factors (e.g., demographic factors) are driving decisions. These factors might be large enough to overwhelm any effects of the experimental manipulation. Yet, in a testament to its robustness as an intervention, a large effect of blinded expertise emerges across both experiments (observed odds ratios were 2.18 and 2.70).
We also closely replicated the credibility rating. As illustrated in Figure 8, the plaintiff’s expert enjoyed boosts in both perceived skill and genuineness (and thus overall credibility). In addition, and perhaps less obviously, we again see a sort of specific contrast effect: not only does the plaintiff’s credibility improve, but the perceptions of the defendant’s genuineness (but not skill) drop (see Robertson & Yokum, 2012, p. 788).

Figure 8. Comparison of credibility ratings between mTurk results reported in Robertson and Yokum (Robertson & Yokum, 2012) and the Tucson reported herein.

The continuous response measurement (CRM) data illustrates a new, useful tool for jury research (see also Yokum & Robertson, in prep, for a more extended discussion of CRM). The data are mostly exploratory, flagging topics for further consideration and
issues for further work. But we also see here its value added when coupled with an experimental design.

In particular, several important insights emerge that could not be uncovered by the non-CRM data. Consider those jurors in the treatment condition who nonetheless found for the plaintiff. Were they perhaps confused by the concept of blind expertise, never fully appreciating its potential; or did they understand but were simply not convinced it was helpful? Either is a plausible reason for why a juror would be unmoved by the blinded expert (in comparison to the handpicked expert). The CRM patterns in Figure 3 cuts against the first explanation of lack of understanding. The mean ratings for eventual plaintiff and eventual defense verdict groups (i.e., the blue lines) march in lock-step during the direct examination wherein the fact of blinded employment is revealed. Apparently both groups understand the concept, at least to the same degree as it translates into a rating bump. The split in verdict amongst those in the treatment group instead comes later, at box B, when the issue of patient responsibility is raised. For a certain segment of the treatment group, this dialogue deflates, so to speak, all of the persuasive ground that had previously been earned by the blinded concept.

We also see evidence here that blinded expertise exerts it persuasive effect via two different pathways. The first is that the plaintiff enjoys a raw credibility boost, so to speak, for merely bringing a blinded expert to the stand. We speculated in Robertson and Yokum (2012) that a litigant might enjoy a sort of halo effect, in the sense that a juror might infer that a party who takes extra pains to select a fair expert—who is so dedicated to avoiding bias—is a party to be trusted more generally. Here we see
significant boosts in the verdict rating for the plaintiff during trial segments where blind expertise per se is being discussed. This is not testimony about the underlying facts of the case, and thus does not reflect any new evidentiary fact of relevance to whether medical negligence occurred. On the other hand, it might reflect a re-weighting of prior evidence.

The other effect, perhaps the more straightforward one, is that jurors weigh more heavily the downstream testimony of the expert. In other words, once they appreciate the expert is blinded (and therefore unbiased), his opinions become more trustworthy. He is more credible. An interesting way this surfaces is as a sort of “immunization” effect. In particular, in the control condition, the defendant wins significant rating ground when crossing the plaintiff’s expert about the possibility of drug seeking. That same cross, however, is not significantly effective whenever the plaintiff’s expert is a blinded expert. The blind apparently provides some immunity from the defense attorney’s attack.

Limitations

Attempting to make comparisons across experimental conditions has revealed that the $60\text{MA} \pm (60\text{SD}\times2)$ statistical rule, although helpful for within-subjects analyses, is perhaps not as useful for between-subjects experimental analyses. The possibility of “masking” (see p. 259) is one problem; the same sized effect might be missed if the manipulation happens to have caused a larger perturbation in the standard deviation of the prior 60 seconds. The second, and more serious problem, is that looking for whether the mean ratings between two groups are on the same side of the $p$ value
fence is really not what we want. That tactic can lead to arbitrarily small differences being flagged. Arguably my “immunization” example falls prey to this problem. If one looks at minute 21, there is a large drop for both control and treatment groups when drug-seeking is insinuated. But only in the control group does the drop surpass the Bollinger band.

An alternative approach is to examine changes in slopes across experimental groups. For example, compare five second intervals between the control and treatment group. For each interval, plot a distribution of the within-subject change in rating. Then you compare the two distributions using typical statistical tools, such as a t-test if normally distributed or otherwise a non-parametric test (e.g., Wilcoxon rank sum). Future work could further develop this possibility.

References


doi:10.1038/485298a
CHAPTER 7: THE NOVEL NEW JERSEY EYEWITNESS INSTRUCTION INDUCES SKEPTICISM
BUT NOT SENSITIVITY

Author Note

This work was conducted in collaboration with Athan P. Papailiou and
Christopher T. Robertson
Abstract

This experiment tested the efficacy of the recently promulgated New Jersey jury instruction, which aims to mitigate jurors’ undue reliance on faulty eyewitness testimony by teaching, in plain language, specific scientific findings about the frailties of human memory. In a 2×2 between-subjects design, mock jurors (N = 335) watched a 35-minute murder trial, wherein identification quality was either “weak” or “strong” and either the New Jersey or a “standard” instruction was delivered. Jurors were more than twice as likely to convict when the standard instruction was used (OR = 2.55; 95% CI = 1.37 – 4.89, p < .001). However, the New Jersey instruction did not improve juror’s ability to discern quality; rather, they indiscriminately discounted “weak” and “strong” testimony in equal measure.

Key words: jury instruction, State v. Henderson, eyewitness testimony, diagnosticity
Background

The United States Supreme Court, about forty years ago, laid out a framework for evaluating the admissibility of eyewitness identification evidence (Neil v. Biggers, 1972, pp. 199–200; see also Manson v. Braithwaite, 1977). A catalog of empirical studies since then has proven that admissible eyewitness identifications, especially when extracted from suggestive lineup procedures, can nonetheless be surprisingly misleading to jurors (see Nadel & Sinnott-Armstrong, 2012, Part Two, pp. 29–160, for a review). As the list of defendants wrongfully convicted due to eyewitness testimony continues to grow, the legal system has begun searching in earnest for a reform solution. One category of reform is to revisit the admissibility rules themselves, in an effort to more aggressively police what testimony reaches the jury. A second category of reform instead focuses on jurors as consumers of the (already admitted) testimony. The aim here is to assist jurors in distinguishing trustworthy from unreliable testimony, by teaching them about the foibles of human memory, especially when tainted by unduly suggestive lineup procedures. As advocates of this strategy have noted, “we now have enough empirical evidence . . . to insist that jurors should be informed about the proneness to error of whatever [identification] procedure is used” (Laudan, 2012, p. 274).

The New Jersey judiciary has recently taken the mantle of this second reform movement, and in July of 2012 announced the release of a new judicial instruction, one carefully constructed to inform lay jurors of the state-of-the-science on eyewitness memory and how to leverage that knowledge in assessing such testimony (New Jersey
Courts, 2012; see also Report of the Special Master, State v. Henderson, No. A-8-08, 2011). The instruction admonishes jurors that “[e]yewitness identification evidence must be scrutinized carefully,” since “research has shown that there are risks of making mistaken identifications” (New Jersey Eyewitness Instruction, 2012). It proceeds to dispel the belief that memory is “like a video recording,” and instead explains, in layperson terms, that the memory process is composed of acquisition, retention, and retrieval stages. The core of the instruction, however, concerns “specific factors that [the juror] should consider,” generally pertaining to “the observations and perceptions on which the identification was based, the witness’s ability to make those observations and perceive events, and the circumstances under which the identification was made.” Guidance is given on specific factors, for example, advice is given that “a witness’s level of confidence, standing alone, may not be an indication of the reliability of the identification,” and that an officer administering a lineup, if not blinded as to who is the suspect, “may intentionally or unintentionally convey that knowledge to the witness.” (These factors are further discussed in the Methodology section below, in terms of their relevance to the trial stimulus reviewed by our mock jurors, and the full instruction on eyewitness testimony can be found in Attachment E.)

The New Jersey instruction has received considerable advance praise. Distinguished psychologists in the areas of memory and eyewitness testimony have applauded the New Jersey instruction in that it “relie[s] on, and receives strong support from, decades of research from cognitive psychology” (Schacter & Loftus, 2013, p. 120). New Jersey Supreme Court Chief Justice Stuart Rabner noted that “[t]he instructions are
7. The novel NJ instruction

designed to minimize the risk of wrongful convictions and help jurors reach informed, just decisions” (New Jersey Courts, 2012). A press release from the Innocence Project commended the instruction as “the first in the nation that explain[s] the way memory works and the factors that can affect the reliability of eyewitness identifications” (Innocence Project, 2012). Director Barry Scheck proclaimed that “these instructions will revolutionize the way that juries scrutinize identification evidence” (Innocence Project, 2012).

We too applaud the New Jersey initiative for its serious approach to summarizing the scientific literature, but we ask a different question – an empirical question – namely: Does administration of an enhanced judicial instruction, such as the one promulgated by New Jersey, actually affect, and ideally improve the sensitivity to quality, of jury decision making? The randomized controlled experiment reported herein provides initial insights into these questions, and indicates that, unfortunately, the new instructions might not be as efficacious as hoped.

A uniquely new instruction

To our knowledge, no prior empirical study has tested the New Jersey instruction. Past attempts to improve the efficacy of other textual instructions on eyewitness instruction, however, have generally not succeeded as hoped. A notable example is the Telfaire instruction, adopted during the 1970s and now the most commonly given cautionary instruction (Penrod & Cutler, 1995). Several experimental tests have shown that the Telfaire instruction fails to improve sensitivity to the quality of the eyewitness testimony. A grab-bag of mixed results have instead found everything
else: no effect (Cutler, Dexter, & Penrod, 1990), across-the-board skepticism (Greene, 1988), and perhaps even across-the-board credulity (Zemba & Geiselman, reported as under review in Penrod & Cutler, 1995, but apparently never published). The first experiment by Greene (1988), for example, used a 2 × 2 between-subjects design with a videotaped mock criminal trial, wherein the identification was either strong or weak and the Telfaire instruction either was or was not provided. The instruction pushed guilty rates to the floor, reducing guilty verdicts from 41% to 7% in the strong condition, and pressing flat at 3% across the weak conditions – functionally then, it actually desensitized jurors, rendering them skeptical of the otherwise strong testimony. A second experiment found that a revised Telfaire instruction – one with simplified, reorganized language – again induced skepticism without any sensitivity gains. Work by Cutler, Dexter, & Penrod failed to replicate a skepticism effect, but again observed that “[t]he Telfaire instruction proved completely ineffective at sensitizing jurors to eyewitness evidence” (1990, p. 1205). A review by Penrod & Cutler (1995, p. 835) lamented that “judges' instructions do not serve as an effective safeguard against mistaken identifications and conviction – they may produce a slight skepticism effect, but they appear to reduce sensitivity to witnessing and identification conditions.”

The New Jersey instruction, however, is notably different than the Telfaire instruction. It provides substantially more detail about memory mechanisms and how an eyewitness identification can be misleading. Such additional instruction has been found effective, albeit in a different pedagogical format. Pawlenko, Safer, Wise, & Holfeld (2013), for example, tested whether a 27-slide PowerPoint presentation (which
they called the “interview-identification-eye teaching aid,” or the “I-I-Eye aid” for short), which explained various factors affecting the reliability of eyewitness identifications, affected verdict rates amongst mock jurors ($N = 293$ undergraduate students) reading a vignette of a criminal murder trial. They found a promising result, namely, that the I-I-Eye aid increased sensitivity to the strength of identification procedures, relative to two different standard instructions. Guilty verdict rates, when standard instructions were read, floated around 30% regardless of whether the eyewitness testimony was weak or strong. With the I-I-Eye aid provided, in contrast, the guilty verdict rates were 55% or 15%, respectively, depending on if the testimony was strong or weak – a significant increase in sensitivity.

Despite the promising result found by Pawlenko and colleagues, it might be difficult to generalize to, and implement within, the real court setting. It is, for example, unclear how the I-I-Eye aid would actually be delivered, for example whether by the judge or via expert testimony. Courts might be more likely to implement a more traditional textual instruction, read aloud to the jury, than to venture into the more uncharted waters of classroom instructor and PowerPoint presenter. To this extent, research is required into the efficacy of enhanced instruction as presented verbally from a judge.

---

55 A second concern – of external validity rather than practicality – is that the PowerPoint stimulus might have been relatively engaging in comparison to a read transcript of a trial, while it would be relatively mundane in comparison to the live action of the courtroom. (The Pawlenko task entailed mixed media: a written vignette of the trial, but a PowerPoint slide show of the I-I-Eye Aid).
Our study thus bears similarity to Pawlenko et al.’s study, at least in terms of its aim, but ultimately we are concerned with the New Jersey instruction on its own merits, as an exemplar of how read-aloud instructions from the judge might be revised to more effectively teach about the foibles of eyewitness testimony. And it is at this empirical nexus that the New Jersey instruction holds out its promise. Scholars have speculated that “it might be necessary to educate jurors about the fallibility of identification witnesses in more detail” (Bornstein & Hamm, 2012, p. 53). The New Jersey instruction aims to do precisely that. Will the enhanced instruction mimic the efficacy of the I-I-Eye aid, by improving sensitivity, or will it instead continue to flounder like the Telfaire instruction? This experiment provides data to begin answering that question.

Method

Design and stimuli

A 2 × 2 between-subjects factorial design was used, wherein participants acted as mock jurors by watching one of four videos of a criminal trial, systematically varied (through video editing) by the strength of eyewitness testimony (“ID Quality”: weak or strong) and type of judicial instruction used (“Instruction”: standard or enhanced). The video, although abbreviated (it lasted from 30 – 40 minutes depending on condition), contained a relatively rich set of trial elements: opening statements from both the prosecution and defense; direct and cross-examination of three witnesses; closing arguments; and jury instructions read aloud by the judge.56

56 Videos and transcripts are available upon request from the authors.
The trial, *State of New Jersey v. Peter Brown*, involved allegations that the defendant robbed and murdered the cashier of a gas station convenience store.\textsuperscript{57} The case hinged on the testimony of a witness who was present when the crime happened, and who identified the defendant as the culprit during a photographic lineup.

Video editing was used to manipulate the strength of this eyewitness testimony, according to several of the criteria delineated as important within the novel New Jersey instruction (and by extension the body of empirical research on eyewitness testimony). For example, in the “strong” ID Quality condition, the officer administering the lineup was blinded as to which lineup participant was the suspect, the witness was instructed that the suspect may or may not be in the lineup, and the lineup included eight photographs. In the “weak” ID Quality condition, in contrast, the officer was aware of who was the suspect, failed to instruct that the suspect may be absent, and used only five photographs. Table 1 provides a full list of the differences across ID Quality conditions. All other aspects of the testimony – and indeed all other evidence – were held constant across conditions.

\textsuperscript{57} The video was an enactment, with actors, of the vignette transcript used by Nell Pawlenko et al. (2013). Our thanks to those authors for sharing their materials, and allowing us to develop it into a video format.
Table 1. Operationalization of ID Quality. Columns indicate differences in the key witness's testimony across the “weak” and “strong” ID Quality conditions. All other aspects of the evidence were held constant across conditions.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Weak</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the interviewing officer ask the eyewitness about media exposure?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Was the eyewitness instructed to avoid discussing the crime and avoid the media?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Did the interviewing officer ask leading questions about the appearance of the perpetrator?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Did the interviewing officer ask leading questions about the quality of the witness's view of the perpetrator?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Were standardized identification procedure instructions used?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>How many photos were in the lineup array?</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Did the description of the suspect match the appearance of line-up participants?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Was the interviewing officer unaware of which lineup participant was the suspect?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Was the witness instructed that the perpetrator may or may not be in the lineup?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Did the interviewing officer provide confirmatory feedback immediately after the identification?</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

With this operationalization of ID Quality, jurors should – if they are accurately assessing the quality of eyewitness testimony – be significantly more likely to convict the defendant in the “strong” condition than in the “weak” condition. Indeed, the two ID Quality conditions were deliberately crafted to embody obviously different indicia of eyewitness quality, thereby providing fertile ground for the New Jersey instruction to do its salutary work, if in fact it does so work.
The Instruction variable was manipulated by having the judge use either a “standard” instruction, modeled after the instruction commonly used in Florida, or the novel instruction recently promulgated by the New Jersey courts (*New Jersey Eyewitness Instruction*, 2012) (see also Attachment E), which we label as the “enhanced” instruction for convenience. Florida was selected, as opposed to the older New Jersey instruction, in order to assess how the “enhanced” instruction stacks up against other language currently in use. (This was, for example, the instruction used in the recent, high-profile case of Trayvon Martin). Moreover, the Florida instruction is relatively sparse, without providing much precise guidance; it instead contains general questions to consider, such as “Did the witness seem to have an accurate memory” or “Did the witness seem to have an opportunity to see and know the things about which the witness testified?” It does not teach about the frailty of human memory or the potentially biasing impact of unduly suggestive lineup procedures. In other words, it lacks precisely the sort of empirically informed facts that the novel New Jersey instruction sought to warn about. (The text of the “standard” instruction, as it pertains to eyewitness testimony, is found in Attachment F.) Thus, if the unique instructional elements of the New Jersey

---

58 See Florida Standard Jury Instructions (2012), Rule 3.9, Weighing the Evidence. This was, for example, the instruction used in the recent, high-profile case of Trayvon Martin. See *Zimmerman final jury instructions*, L.A. TIMES.COM, [http://documents.latimes.com/zimmerman-final-jury-instructions](http://documents.latimes.com/zimmerman-final-jury-instructions) (page 17) (July 12, 2013). The exact language used is that found within Pawlenko et al. (2013), from which we adapted the trial transcript into a video.
instruction have a salutary impact on how juries assess eyewitness testimony, then it should be especially notable in comparison to the “standard” instruction.\textsuperscript{59}

Crossing the ID Quality and Instruction variables affords an opportunity to test two aspects of the New Jersey instruction. First, does it have any effect on jury decision-making? A finding that jurors are less (or more) likely to find the defendant guilty when the “enhanced” instruction is used relative to the “standard” instruction would provide an affirmative answer. Second, and importantly, if there is an effect, is it an effect that increases diagnostic accuracy, that is, sensitivity to eyewitness testimony quality? This would be revealed by an interaction effect: ideally, the use of the “enhanced” instruction should result in lower conviction rates when the ID Quality is “low” rather than “high.” After all, if the “enhanced” instruction similarly decreased convictions in the “high” ID Quality condition, then it would obtain a decrease in false positives, but only at the cost of an increase in false negatives.\textsuperscript{60} To fully test this possibility, the case facts (across ID Quality conditions; see Table 1) were deliberately manipulated to either avoid or succumb to risks explicitly laid out in the “enhanced” (but not “standard”) instructions. For example, the lineups in the “weak” and “strong”

\textsuperscript{59} Both conditions also contained the same additional instruction, unrelated to eyewitness testimony, found in Attachment E. Note that, due to a programming error, this supplementary instruction appeared near the beginning of the “enhanced” instruction but at the end of the “standard” instruction. Strictly speaking this is a confound, but it seems unlikely to have caused any effect, especially given the brevity of the “standard” instruction.

\textsuperscript{60} This is not necessarily a bad consequence relative to the status quo; as Blackstone noted, it is “better that ten guilty persons escape than that one innocent suffer.” COMMENTARIES 358. New Jersey Chief Justice Rabner’s articulated goal for the instruction, quoted in the introduction (i.e., “[t]he instructions are designed to minimize the risk of wrongful convictions”), suggests that he would be satisfied on this front. The point here is merely that it would be better still if the instruction reduced false positives without increasing false negatives.
conditions entailed five and eight photos, respectively, while the “enhanced” instruction directs that “[a] minimum of six persons or photos should be included.” (The “standard” instruction is silent on the issue). Thus, a juror following the “enhanced” instruction should discount the testimony in the “weak” condition, but not in the “strong” condition.

Participants

The study population was drawn from Amazon Mechanical Turk, with an advertisement to participate in an approximately one-hour research experiment in exchange for $2.00 payment. Five hundred and four persons proceeded past the informed consent page, although 143 cases (28%) were removed for failure to complete the task (most left during the video), and another 26 cases (5%) were removed for one of two quality-assurance reasons: (a) entry of “garbage” text (e.g., smashed keys or copied-and-pasted question prompts); or (b) completing the experiment in less than 30 minutes, an impossibly fast time (since the video stimuli lasted at least 30 minutes). The final sample (N = 335) was predominantly white (80%), tended to be female (57%), and was about 35 years old. Most (83%) had at least some college credit.

61 See Paolacci, Chandler, & Ipeirotis (2010) for an overview of Amazon Mechanical Turk and the demographic composition of its users relative to other online and offline sources.

62 The precise demographics were as follows: the racial composition was 80% White, 10% Black or African American, 0.1% American Indian or Alaskan Native, 5% Asian, 0.2% Native Hawaiian or Other Pacific Islander, and 4% “some other race”; the mean, median, and standard deviation of age were, respectively, 36, 33, and 12; and the breakdown of highest educational obtainment was 0.0% Elementary School, 0.2% Some High School, 14% High School Graduate, 30% Some College Credit, 13% Associate Degree, 36% Bachelor’s Degree, 5% Master’s Degree, 1% Professional Degree, and 1% Doctoral Degree.
Procedure and measured variables

Participants completed the experiment online, via a survey programmed and hosted on Qualtrics (see http://www.qualtrics.com/). The survey flow entailed an initial informed consent page, a demographics questionnaire, and then an instruction page including the admonition to “please treat this case with the same seriousness and diligence that you would use in a real trial.” The experiment itself followed: participants watched one of the four mock trial videos, and then answered a series of questions about their verdict preferences and use of the evidence.

The primary dependent variable of interest, answered immediately after the video, was a binary response (“Yes, guilty” or “No, not guilty”) to the question, “Based on the evidence and the instructions provided by the judge, has the prosecutor proved, beyond a reasonable doubt, that the Defendant is guilty?” A follow-up question asked for a brief written description of the reasons for their decision.

Additional questions probed issues of judicial instruction comprehension, confidence in verdict, use of evidence, and the reliability of the eyewitness testimony. Responses were indicated on 6-point Likert scales. Specifically, participants were asked to assess the following (with Likert anchors in parentheses): “I understand the Judge's instructions” (“strongly disagree” to “strongly agree”); “How confident are you in your verdict?” (“not at all confident” to “absolutely confident”); and “Barbara Dunn's eyewitness testimony was influential in your verdict decision” (“strongly disagree” to “strongly agree”).
Results

Verdict

Figure 1 shows the proportions of guilty verdicts by experimental condition. As can be seen, the “enhanced” instruction reduced the number of guilty verdicts, although it appears to do so about equally for both levels of ID Quality. In the “standard” Instruction condition, conviction rates are similar in the “weak” (23%) and “strong” (26%) ID Quality conditions, and remain similar – albeit substantially reduced – when the “enhanced” instruction is used (weak = 9%; strong = 12%).
Figure 1. Proportion of Guilty verdicts (and 95% confidence intervals) by Instruction and ID Quality.

Binary logistic regression was used to statistically assess the impact of ID Quality and Instruction (and their interaction) on the likelihood of a guilty verdict, as well as provide additional control (beyond randomization) of demographic covariates.
Table 2 displays the results of three regression models. The first model regressed verdict (0 = Not Guilty; 1 = Guilty) on the independent variables using dummy coding: Instruction_Standard (0 = enhanced; 1 = standard) and ID Quality_Weak (0 = strong; 1 = weak). Model two builds upon model one, by adding the possible Instruction by ID Quality interaction term. Model three includes demographic covariates of Male (0 = female; 1 = male), Minority (0 = white; 1 = non-white), College (0 = less than Bachelor's degree; 1 = Bachelor's degree or higher), and Age_10 (age at 10 year intervals); since the interaction term was non-significant even with demographic covariates, and its inclusion complicates interpretation of the other coefficients, model three does not include it.
Table 2. Logistic regressions predicting a verdict of guilt.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (S.E.)</td>
<td>β (S.E.)</td>
<td>β (S.E.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.04 (.29)</td>
<td>-1.99 (.33)</td>
<td>-2.59 (.55)</td>
</tr>
<tr>
<td>Instruction_Standard</td>
<td>1.02 (.31) **</td>
<td>0.95 (.42) *</td>
<td>0.94 (.32) **</td>
</tr>
<tr>
<td>ID Quality_Weak</td>
<td>-0.24 (.31)</td>
<td>-0.69 (.49)</td>
<td>-0.22 (.29)</td>
</tr>
<tr>
<td>Instruction_Standard × ID Quality_Weak</td>
<td>-</td>
<td>0.17 (.63)</td>
<td>-</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>-</td>
<td>-0.13 (.31)</td>
</tr>
<tr>
<td>Minority</td>
<td>-</td>
<td>-</td>
<td>0.65 (.35) *</td>
</tr>
<tr>
<td>College</td>
<td>-</td>
<td>-</td>
<td>0.72 (.32) *</td>
</tr>
<tr>
<td>Age_10</td>
<td>-</td>
<td>-</td>
<td>0.02 (.12)</td>
</tr>
</tbody>
</table>

Null deviance (df) 311.86 (334) 311.86 (334) 311.86 (334)
Residual deviance (df) 299.54 (332) 299.47 (311) 290.19 (328)

Note. Null and residual deviances are shown at the bottom of the table, for assessing overall model fit. *p < .10, *p < .05, **p < .01, ***p < .001.

The statistical results confirm what visual inspection of Figure 1 suggests.

Referring to model three (which is the model of best fit, as revealed be the reduction in residual deviance; see Table 2), there is a substantial main effect of Instruction: use of the “standard” instruction relative to the “enhanced” instruction more than doubles the odds that the defendant will be found guilty (odds ratio = 2.55; 95% CI = 1.37 – 4.89; p < .001). ID Quality, on the other hand, does not have a significant main effect on verdict.
The novel NJ instruction \((p = .44)\). There is also no evidence of an interaction effect \((p = .84)\). To summarize it differently, use of the novel New Jersey instruction substantially reduced the likelihood that the defendant would be found guilty, but its reducing effect was the same regardless of whether the eyewitness identification testimony was weak or strong.

**Self-Perceived comprehension, confidence, and influence**

Table 3 summarizes the means and standard deviations of the Likert-scale responses. Analysis of Variance (ANOVA) was used to assess whether any main effects of Instruction or ID Quality, or the interaction of the two, were statistically significant.

<table>
<thead>
<tr>
<th>ID Quality</th>
<th>Instruction</th>
<th>(n)</th>
<th>Comprehension of Instruction</th>
<th>Confidence in Verdict</th>
<th>Influence of Testimony</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Enhanced</td>
<td>83</td>
<td>5.51 (0.85)</td>
<td>4.61 (1.17)</td>
<td>4.12 (1.38)</td>
</tr>
<tr>
<td></td>
<td>Standard</td>
<td>88</td>
<td>5.61 (0.56)</td>
<td>4.77 (1.13)</td>
<td>4.43 (1.31)</td>
</tr>
<tr>
<td>Weak</td>
<td>Enhanced</td>
<td>80</td>
<td>5.56 (0.61)</td>
<td>4.66 (1.08)</td>
<td>4.15 (1.34)</td>
</tr>
<tr>
<td></td>
<td>Standard</td>
<td>84</td>
<td>5.61 (0.62)</td>
<td>4.65 (1.19)</td>
<td>4.40 (1.14)</td>
</tr>
</tbody>
</table>

*Note.* Means (and standard deviations) of mock jurors' comprehension of the judicial instruction, confidence in their verdict, and agreement that the testimony of the key eyewitness was “influential in [their] verdict decision.” All responses were on a 1-6 point Likert scale.

Self-perceived comprehension of judicial instructions was not significantly affected by either ID Quality \((p = .74)\) or Instruction \((p = .29)\), and no interaction was

---

63 The interaction is not significant in model two nor a full model (such as model three) containing demographic covariates; the \(p = .84\) value is from the full model.
found ($p = .67$). Responses instead hovered around 5.55 across conditions, indicating that most subjects “somewhat agree[d]” or “agree[d]” that they understood the judge’s instructions, regardless of instruction type.

Confidence in verdict was likewise unaffected by ID Quality ($p = .76$), Instruction ($p = .54$), or the interaction of the two ($p = .51$). These scores were all within the 4.6 – 4.7 range, indicating most jurors were “confident” in their verdicts.

There was, however, an effect of Instruction on self-assessments of whether the testimony of the key eyewitness influenced participants’ verdicts. In particular, mock jurors who received the “enhanced” instruction were significantly more likely to believe (correctly) that the eyewitness testimony did *not* influence their verdict ($F(1, 331) = 4.00, p = .046$).64 This was true in both the “strong” and “weak” ID Quality conditions, where estimates of influence were about 0.30 points (on the 6-point Likert scale) lower when the “enhanced” instruction was used (specifically, 4.12 versus 4.43 in the “strong” condition, and 4.15 versus 4.40 in the “weak” condition). Lastly, the effect of ID Quality was again non-significant ($p = .99$), as was the interaction term ($p = .84$).

**Discussion**

This study tested the efficacy of one approach to ensuring that jurors have decreased reliance on faulty testimony and increased reliance on trustworthy testimony. Namely, provide them with an “enhanced” judicial instruction, one that cautions against undue reliance on imperfect human memory, especially in the presence

---

64 This belief was “correct” in the sense that the referenced testimony was incriminating, and yet those in the “enhanced condition” were not more likely to convict.
of suggestive lineup procedures. The New Jersey judiciary, after a careful review of the empirical evidence on eyewitness testimony, crafted precisely such an instruction. It could serve as a model for other jurisdictions – if, that is, it succeeds in its task.

Unfortunately, our experimental data provides a mixed review of the efficacy of the New Jersey instruction. On the one hand, the use of the New Jersey instruction did, as would be hoped, substantially reduce juror reliance on weak identification evidence, as compared to the Florida-based instruction. This suggests that jurors can and do in fact listen to the judge's admonitions, and they take into consideration that guidance. This is an especially promising result when considered against the backdrop of empirical work often finding that jurors are unaffected by instructions. It is, perhaps, partly a testament to the admirable job that the New Jersey judiciary did in ensuring that the language of its instruction was clear and easy to understand. When jurors fail to follow instructions, it is often because they do not understand what is being asked of them (see generally, (Devine, 2012, pp. 55–56). Here, as evidenced by the high ratings of self-perceived comprehension of judicial instructions (most jurors at least “somewhat agree[d]” they understood), it seems they did not find themselves so baffled. Of course, the fact that jurors self-report understanding does not necessarily mean they are not confused. The review by Devine (2012, p. 56.) notes the “disconcerting finding” that, despite poor performance on comprehension tests, “jurors [nonetheless] typically

---

Nietzel, McCarthy, & Kerr (1999), for example, in a meta-analysis of 150 relevant effect sizes, find only a weak correlation between conditions of instruction and no-instruction and effect on juror behavior (mean $r = 0.07$)
report understanding the judge's instructions with no problems.” This could not be the full story regardless, since the language of the Florida instruction was also not perceived as particularly complex; our mock jurors reported understanding its language about equally well. (More on this in a moment.)

On the other hand, the New Jersey instruction also reduced, in equal amount, juror reliance on strong identification evidence. The failure to find an interaction of the enhanced instruction with the quality of the eyewitness testimony undermines the belief that the New Jersey instruction increases sensitivity. It seems instead that jurors, confronted with a catalog of the foibles of human memory and the extra risks posed by unduly suggestive lineup procedures, compensated – not by selectively applying the instructed information to the facts at hand – but rather by indiscriminately discounting any and all eyewitness identification testimony. Needless to say, this is a suboptimal result, one that increases the rate at which a guilty defendant is nonetheless exonerated (i.e., a false negative). Yet it might still be an improvement over the “standard” instruction, at least if one agrees with Blackstone's argument that reducing false positives is more important than reducing false negatives (“better that ten guilty persons escape than that one innocent suffer.”).

What could account for the across-the-board skepticism effect? One possibility is that although jurors are swayed by the content of the instruction (in this case the empirically informed details about how to scrutinize eyewitness testimony) and want to apply this knowledge to the case at hand, they cannot, for whatever reason, do so accurately. Perhaps, for instance, our mock jurors knew they should discount
testimony whenever a certain criterion was met, but were unsure at what threshold that criterion should be deemed satisfied. The “enhanced” instruction is often silent as to such issues of magnitude. For example, it states that “high levels of stress can reduce an eyewitness’s ability to recall and make an accurate identification,” but says nothing about what constitutes a “high” level. As one more example, it notes that the “amount of time any eyewitness observes an event may affect the reliability of an accurate identification,” but provides little guidance on the duration-to-reliability function, except for noting a difference between “brief or fleeting contact” and “more prolonged exposure.” A juror unsure of whether a threshold has been met might, as the law instructs (in terms of reasonable doubt), err on the side of doubt and presume the testimony, even if otherwise “strong,” surpasses the threshold. This could explain why the enhanced disclosure reduces reliance on even the “strong” testimony. On the other hand, the “weak” versus “strong” conditions used here were (deliberately) operationalized along criteria that did, in fact, have relatively bright lines (e.g., either the interviewing officer was blinded or not; either standardized instructions were used or not; either six or more participants were used in the lineup or not; etc…). Thus an inability to apply fuzzy criteria seems at best a partial explanation.

An alternative explanation, which our present data cannot rule out, is that it is not the content, but rather the mere length of the “enhanced” instruction, that is driving the observed effect. The “standard” instruction took about five minutes to read, whereas the “enhanced” instruction took about 15 minutes to read. This three-fold increase in exposure to talk about the problems of eyewitness testimony could make
the issue more salient and, therefore, more of a concern in general, whatever the case facts may be. The design of our experiment cannot distinguish this possibility from a failure to accurately apply the criteria (a limitation of internal validity). But the New Jersey instruction, relative to other extant instructions, simply is that much longer. Thus, for purposes of external validity, and in particular for appreciating the likely effect of the New Jersey instruction, the ultimate effect of interest might be the same regardless of underlying reason: jurors are substantially less likely to convict on the basis on eyewitness testimony. Nonetheless, future experiments should tease apart the role of length versus ability to apply the guidelines, the results of which would be useful for further improving the judicial instruction.

An additional explanation, speculative but perhaps holding more promise, is that the observed skepticism effect reflects more about source credibility (and by extension the depth of cognitive processing) than the effectiveness of the instruction per se. In particular, jurors might interpret the judge’s decision to read the enhanced instruction — one which belabors the inaccuracies of eyewitness testimony, without much counterbalanced discussion of its conditions of accuracy — as an implicit signal that said testimony should, in fact, not be trusted. Why else, a juror might wonder, would the judge bother to make such a fuss about it?

---

66 On the other hand, a real trial would be substantially longer than that of our mock trial stimuli. Perhaps the effect of instruction length is dependent on what proportion of the total trial experience it occupies. In our study, 5 and 15 minutes of instruction in an approximately 35 minute trial constitutes 14% and 43% of the total trial — a substantial difference. In a 25 hour trial, in contrast, the proportions would only constitute 0.3% and 1%, a difference which is likely to be lost relative to the bulk of the trial experience.
A related dynamic has been described by Cutler et al. (1990), who found that adversarial expert testimony induced greater sensitivity but that non-adversarial expert testimony (viz., a court appointed expert) resulted in a skepticism effect with no sensitization. They proposed an explanation in terms of the elaboration likelihood model of persuasion (ELM; see generally (Petty & Cacioppo, 1986)). The ELM predicts that people (not unreasonably) process and scrutinize information more critically when the source of that information is perceived as less credible \textit{a priori}. With a credible source, in contrast, statements are more likely to be accepted on face value. The link for present purposes is that a juror, who trusts the judge and misinterprets the judge’s lengthy instruction as an implicit signal the testimony is faulty, might fail to engage in the increased cognitive elaboration required to apply the instruction to the facts at hand. Easier instead to simply discount the testimony wholesale, without any elaboration.

A source credibility explanation could also accommodate the fact that the I-I-Eye Aid used by Pawlenko et al. (2013) induced skepticism. Specifically, the I-I-Eye Aid was \textit{not} presented within the context of the mock trial. Study participants rather viewed the PowerPoint slides before engaging the trial materials, in a manner suggesting it was the researchers, not the judge, providing the guidance. This is not a context wherein subjects would misinterpret the guidance as an implicit signal from a court.

The present study is not designed to test source credibility effects. The relevant test, of course, would be to have the content of the New Jersey instruction delivered by someone other than the judge. Who that might be in practice is a complicated issue,
but methodologically in the laboratory, one could simply pull the instruction outside the context of the mock trial, similar to how Pawlenko et al. (2013) presented the I-I-Eye Aid; the prediction being that in such case the instruction would induce sensitivity. (Likewise, one would predict that the I-I-Eye Aid, if presented by the judge rather than researchers, could induce skepticism rather than sensitivity).

As a mock jury study with an online study population, the usual cautions apply. The case was not a full-length trial, was hypothetical rather than placing real life and limb in danger, used a sample that was not precisely representative of the jury pool, and so forth. Nonetheless, this experiment was a randomized, controlled trial with several hundred jury eligible adults. And it used a 30 – 40 minute trial stimulus of relatively high ecological validity, in terms of providing a video capture of all key trial elements, and verbatim use of the New Jersey instruction. The lackluster performance of the New Jersey instruction should thus signal that the task of reforming judicial instructions on eyewitness testimony is not complete. Our results provide a warning to judges to be wary of the possibility that enhanced instructions will lead jurors to overly discount reliable testimony. Our results also motivate the need for yet still further research on how judicial instructions on eyewitness testimony can be best crafted, and how they can be best delivered.

We end with one final point to begin considering. Even if the “enhanced” instruction caused jurors to indiscriminately discount testimony, it might still be leveraged – by way of a more active judicial role – to ultimately increase the diagnosticity of trials. In particular, one suggestion is to allow judges discretion as to
when to apply the instruction, and as to which parts of it to read. Some of this is already done, in obvious fashion: if there was no lineup conducted, then obviously there is no need to give cautionary instructions pertaining to a lineup procedure. But a more aggressive tactic would be, for example, to allow the judge to decline to read the lineup instruction, even if one was performed, so long as the used procedures met certain gold-standard criteria. For instance, if the lineup entailed more than six photographs, the presiding officer was blinded as to who was the suspect, and so forth, then the judge could elect to simply not read the warning. In terms of our experiment, this would be akin to reading the instruction in the “weak” condition but not reading the instruction in the “strong” condition. This could avoid unduly alarming the jurors in the latter scenario, and thereby avoid the increased rate of false negatives that we observed experimentally. Of course, this suggestion does raise its own problems, neither of which is trivial. The first is that judges would then be tasked with deciding when the “gold-standard” has been met. Can they do this accurately? This is another empirical question. Second, assuming they can be accurate, does such an active role infringe upon the province of the jury? This is a legal issue that would need to be worked out.
Mr. Peter Brown, as part of his general denial of guilt, contends that the State has not presented sufficient reliable evidence to establish beyond a reasonable doubt that he is the person who committed the alleged offenses. The burden of proving the identity of the person who committed the crime is upon the State. For you to find this defendant guilty, the State must prove beyond a reasonable doubt that this defendant is the person who committed the crime. The defendant has neither the burden nor the duty to show that the crime, if committed, was committed by someone else, or to prove the identity of that other person. You must determine, therefore, not only whether the State has proven each and every element of the offenses charged beyond a reasonable doubt, but also whether the State has proven beyond a reasonable doubt that this defendant is the person who committed it.

The State has presented the testimony of Mrs. Barbara Dunn. You will recall that this witness identified the defendant in court as the person who committed the crime of armed robbery. The State also presented testimony that on a prior occasion before this trial, this witness identified the defendant as the person who committed these offenses. According to the witness, her identification of the defendant was based upon the observations and perceptions that she made of the perpetrator at the time these offenses were being committed. It is your function to determine whether the witness’s identification of the defendant is reliable and believable, or whether it is based on a mistake or for any reason is not worthy of belief. You must decide whether it is sufficiently reliable evidence that this defendant is the person who committed these offenses charged.

Eyewitness identification evidence must be scrutinized carefully. Human beings have the ability to recognize other people from past experiences and to identify them at a later time, but research has shown that there are risks of making mistaken identifications. That research has focused on the nature of memory and the factors that affect the reliability of eyewitness identifications.

Human memory is not foolproof. Research has revealed that human memory is not like a video recording that a witness need only replay to remember what happened. Memory is far more complex. The process of remembering consists of three stages: acquisition – the perception of the original event; retention – the period of time that passes between the event and the eventual recollection of a piece of information; and retrieval – the stage
during which a person recalls stored information. At each of these stages, memory can be affected by a variety of factors.

Relying on some of the research that has been done, I will instruct you on specific factors you should consider in this case in determining whether the eyewitness identification evidence is reliable. In evaluating this identification, you should consider the observations and perceptions on which the identification was based, the witness’s ability to make those observations and perceive events, and the circumstances under which the identification was made. Although nothing may appear more convincing than a witness’s categorical identification of a perpetrator, you must critically analyze such testimony. Such identifications, even if made in good faith, may be mistaken. Therefore, when analyzing such testimony, be advised that a witness’s level of confidence, standing alone, may not be an indication of the reliability of the identification.

If you determine that the out-of-court identification is not reliable, you may still consider the witness’s in-court identification of the defendant if you find that it resulted from the witness’s observations or perceptions of the perpetrator during the commission of these offenses, and that the identification is reliable. If you find that the in-court identification is the product of an impression gained at the out-of-court identification procedure, it should be afforded no weight. The ultimate question of the reliability of both the in-court and out-of-court identifications is for you to decide.

To decide whether the identification testimony is sufficiently reliable evidence to conclude that this defendant is the person who committed these offenses charged, you should evaluate the testimony of the witness in light of the factors for considering credibility that I have already explained to you. In addition, you should consider the following factors that are related to the witness, the alleged perpetrator, and the criminal incident itself. In particular, you should consider:

(1) The Witness’s Opportunity to View and Degree of Attention: In evaluating the reliability of the identification, you should assess the witness’s opportunity to view the person who committed these offenses at the time of these offenses and the witness’s degree of attention to the perpetrator at the time of these offenses. In making this assessment you should consider the following:

(a) Stress: Even under the best viewing conditions, high levels of stress can reduce an eyewitness’s ability to recall and make an accurate identification. Therefore, you should consider a witness’s level of stress and whether that stress, if any, distracted the witness or made it harder for him or her to identify the perpetrator.
(b) **Duration**: The amount of time an eyewitness has to observe an event may affect the reliability of an identification. Although there is no minimum time required to make an accurate identification, a brief or fleeting contact is less likely to produce an accurate identification than a more prolonged exposure to the perpetrator. In addition, time estimates given by witnesses may not always be accurate because witnesses tend to think events lasted longer than they actually did.

(c) **Weapon Focus**: You should consider whether the witness saw a weapon during the incident and the duration of the crime. The presence of a weapon can distract the witness and take the witness’s attention away from the perpetrator’s face. As a result, the presence of a visible weapon may reduce the reliability of a subsequent identification if the crime is of short duration. In considering this factor, you should take into account the duration of the crime because the longer the event, the more time the witness may have to adapt to the presence of the weapon and focus on other details.

(d) **Distance**: A person is easier to identify when close by. The greater the distance between an eyewitness and a perpetrator, the higher the risk of a mistaken identification. In addition, a witness’s estimate of how far he or she was from the perpetrator may not always be accurate because people tend to have difficulty estimating distances.

(e) **Lighting**: Inadequate lighting can reduce the reliability of an identification. You should consider the lighting conditions present at the time of the alleged crime in this case.

(f) **Disguises/Changed Appearance**: The perpetrator’s use of a disguise can affect a witness’s ability both to remember and identify the perpetrator. Disguises like hats, sunglasses, or masks can reduce the accuracy of an identification. Similarly, if facial features are altered between the time of the event and a later identification procedure, the accuracy of the identification may decrease.

(2) **Prior Description of Perpetrator**: Another factor for your consideration is the accuracy of any description the witness gave after observing the incident and before identifying the perpetrator. Facts that may be relevant to this factor include whether the prior description matched the photo or person picked out later, whether the prior description provided details or was just general in nature, and whether the witness’s testimony at trial was consistent with, or different from, her prior description of the perpetrator.
(3) **Confidence and Accuracy**: You heard testimony that Mrs. Barbara Dunn made a statement at the time she identified the defendant from a photo line-up concerning her level of certainty that the photograph she selected is in fact the person who committed the crime. As I explained earlier, a witness’s level of confidence, standing alone, may not be an indication of the reliability of the identification. Although some research has found that highly confident witnesses are more likely to make accurate identifications, eyewitness confidence is generally an unreliable indicator of accuracy.

(4) **Time Elapsed**: Memories fade with time. As a result, delays between the commission of a crime and the time an identification is made can affect the reliability of the identification. In other words, the more time that passes, the greater the possibility that a witness’s memory of a perpetrator will weaken.

In evaluating the reliability of a witness’s identification, you should also consider the circumstances under which any out-of-court identification was made, and whether it was the result of a suggestive procedure. In that regard, you may consider everything that was done or said by law enforcement to the witness during the identification process. You should consider the following factors:

(1) **Line-up Composition**: A suspect should not stand out from other members of the lineup. The reason is simple: an array of look-alikes forces witnesses to examine their memory. In addition, a biased lineup may inflate a witness’s confidence in the identification because the selection process seemed so easy to the witness. It is, of course, for you to determine whether the composition of the lineup had any effect on the reliability of the identification.

(2) **Fillers**: Lineups should include a number of possible choices for the witness, commonly referred to as “fillers.” The greater the number of choices, the more likely the procedure will serve as a reliable test of the witness’s memory. A minimum of six persons or photos should be included in the lineup.

(3) **Multiple Viewings**: When a witness views the same person in more than one identification procedure, it can be difficult to know whether a later identification comes from the witness’s memory of the actual, original event or of an earlier identification procedure. As a result, if a witness views an innocent suspect in multiple identification procedures, the risk of mistaken identification is increased. You may consider whether the witness viewed the suspect multiple times during the identification process and, if so, whether that affected the reliability of the identification.
In determining the reliability of the identification, you should also consider whether the identification procedure was properly conducted.

a. **Double-blind**: A lineup administrator who knows which person or photo in the lineup is the suspect may intentionally or unintentionally convey that knowledge to the witness. That increases the chance that the witness will identify the suspect, even if the suspect is innocent. For that reason, whenever feasible, live lineups and photo arrays should be conducted by an officer who does not know the identity of the suspect. If a police officer who does not know the suspect’s identity is not available, then the officer should not see the photos as the witness looks at them. In this case, it is alleged that the person who presented the lineup knew the identity of the suspect. It is also alleged that the police did/did not compensate for that by conducting a procedure in which the officer did not see the photos as the witness looked at them. You may consider this factor when you consider the circumstances under which the identification was made, and when you evaluate the overall reliability of the identification.

b. **Instructions**: You should consider what was or what was not said to the witness prior to viewing a photo array. Identification procedures should begin with instructions to the witness that the perpetrator may or may not be in the array and that the witness should not feel compelled to make an identification. The failure to give this instruction can increase the risk of misidentification. If you find that the police [did/did not] give this instruction to the witness, you may take this factor into account when evaluating the identification evidence.

c. **Feedback**: Feedback occurs when police officers, or witnesses to an event who are not law enforcement officials, signal to eyewitnesses that they correctly identified the suspect. That confirmation may reduce doubt and engender or produce a false sense of confidence in a witness. Feedback may also falsely enhance a witness’s recollection of the quality of his or her view of an event. It is for you to determine whether or not a witness’s recollection in this case was affected by feedback or whether the recollection instead reflects the witness’s accurate perception of the event.

You may consider whether the witness was exposed to opinions, descriptions, or identifications given by other witnesses, to photographs or newspaper accounts, or to any other information or influence, that may have affected the independence of his/her identification. Such information can affect the independent nature and reliability of a witness’s identification and inflate the witness’s confidence in the identification.
You are also free to consider any other factor based on the evidence or lack of evidence in the case that you consider relevant to your determination whether the identifications were reliable. Keep in mind that the presence of any single factor or combination of factors, however, is not an indication that a particular witness is incorrect. Instead, you may consider the factors that I have discussed as you assess all of the circumstances of the case, including all of the testimony and documentary evidence, in determining whether a particular identification made by a witness is accurate and thus worthy of your consideration as you decide whether the State has met its burden to prove identification beyond a reasonable doubt. If you determine that the in-court or out-of-court identifications resulted from the witness's observations or perceptions of the perpetrator during the commission of these offenses, you may consider that evidence and decide how much weight to give it. If you instead decide that the identification is the product of an impression gained at the in-court and/or out-of-court identification procedures, the identifications should be afforded no weight. The ultimate issue of the trustworthiness of an identification is for you to decide.

If, after consideration of all of the evidence, you determine that the State has not proven beyond a reasonable doubt that Mr. Peter Brown was the person who committed these offenses, then you must find him not guilty. If, on the other hand, after consideration of all of the evidence, you are convinced beyond a reasonable doubt that Mr. Peter Brown was correctly identified, you will then consider whether the State has proven each and every element of these offenses charged beyond a reasonable doubt.
Attachment F: Florida-Based ("Standard") Jury Instructions

It is up to you to decide what evidence is reliable. Some things you should consider are: Did the witness seem to have an opportunity to see and know the things about which the witness testified? Did the witness seem to have an accurate memory? Was the witness honest and straightforward in answering the attorneys’ questions? Did the witness have some interest in how the case should be decided? A juror may believe or disbelieve all of or any part of the evidence or testimony of any witness.
Attachment G: Additional Jury Instructions
[Common to both “Standard” and “Enhanced” Conditions]

Members of the Jury, thank you for your attention. Please listen to the instructions I am about to give you.

Mr. Peter Brown, the Defendant in this case, is accused of first degree felony murder of David Aims. If you have a reasonable doubt as to the guilt of the Defendant you should find the Defendant not guilty. If you have no reasonable doubt you should find the Defendant guilty. It is the evidence introduced at this trial and to it alone that you are to look for that proof. It is up to you to decide what evidence is reliable.

Before you can find the Defendant guilty of the first-degree felony murder, the state must prove the following three elements beyond a reasonable doubt: Number one, David Aims is dead. Number two, did this occur as a consequence of, and while Peter Brown was engaged in the commission of a robbery? Number three, Peter Brown was the person who actually killed David Aims. An issue in this case is whether the Defendant was present when the crime allegedly was committed. If you have a reasonable doubt that the Defendant was present at the scene of the alleged crime, it is your duty to find the Defendant not guilty. Finally, the decision to testify is the 5th Amendment right of the Defendant. The fact that the Defendant in this case did not testify should have no bearing on your verdict.
References


The emergence of applied behavioral science

_Homo economics_, that fictional creature of unlimited information processing power and stable, self-interested preferences, can serve as a useful benchmark for economic modeling—for simplifying efforts to track and predict human behavior, or constructing arguments for how humans should behave in some “rational” framework. Yet for public officials constructing policies and delivering programs to real _Homo sapiens_, the strong neoclassical economic assumptions must often be relaxed in order to achieve the greatest effect by the most efficient means. But how should the assumptions be relaxed, and how can we design interventions that are more sensitive to the quirks of human behavior?

Several fields of research have been merging over the past decades to provide dynamic answers to these important questions. Most notably, researchers in fields across the social and behavioral sciences have, over the past half century, conducted hundreds of experiments to better understand the cognitive processes underlying behavior, such as how people (consciously or, more often, unconsciously) parse complex information, construct and infer their preferences, integrate risk and uncertainty, reason about available options, or convert (or not) intentions into action (for accessible, excellent reviews, see Piattelli-Palmarini, 1994; Kahneman, 2013). Psychologists Daniel Kahneman and Amos Tverky, for example, helped sparked an entire sub-field, judgment and decision-making, dedicated to such topics, and from which a whole catalogue of insights have since emerged (see e.g., Gilovich, Griffin, & Kahneman, 2002; see also Tversky & Kahneman, 1974, for a seminal work). Economists soon began to incorporate
and translate such findings into their reasoning and modeling, and to generate additional findings. An important early article by Richard Thaler (1980), for instance, applied Kahneman and Tversky’s prospect theory (1979) to consumer behavior.

Currently there is a vibrant effort to update policy-making with insights from the behavioral sciences—a body of work best captured under the banner of applied behavioral science (ABS) (see Kahneman, 2012, for an important discussion of why the label “behavioral economics” is not ideal, and advocating instead for “applied behavioral science”). The psychologist Eldar Shafir (2012), for example, collects ABS research in domains varying from employment discrimination to criminal procedure to retirement savings to dietary habits to organ donor registration and beyond (see also DellaVigna, 2007). And perhaps no book captured the imagination of the public—and especially policy makers—like Nudge, by Richard Thaler and Cass Sunstein (2009), wherein the prominent economist and jurist crafted an accessible review of key psychological research and articulated a compelling argument for its critical role in public policy.

The relevance of behavioral science to policy is, of course, not new, nor is its application in select settings (such as disease prevention); but governments are now more explicitly embracing—and expecting their civil servants to use—applied behavioral science, and across the full spectrum of public policy. The U.K. Cabinet Office, for instance, created a Behavioral Insights Team explicitly dedicated to such work and, on the other side of the pond, the White House has assembled its own Social & Behavioral Science Team (SBST); similar initiatives in Israel, Singapore, and elsewhere are
developing. Even before these more salient, coordinated efforts, there was movement
to apply behavioral science in government, especially within the United States. Indeed,
Cass Sunstein, co-author of *Nudge*, served from 2009 to 2012 as Administrator of the
Office of Information and Regulatory Affairs within the Office of Management and
Budget (OMB). Several notable, recent policies have incorporated insights from the
applied behavioral sciences, such as the CARD Act (e.g., requiring disclosure of interest
payments in easier to appreciate formats), § 1511 of the Affordable Care Act (requiring
large employers to *automatically* enroll workers into health insurance), or the redesign
of the motor vehicle fuel economy label (e.g., including easier-to-comprehend gallons
per 100 miles rather than only the nonlinear—and intuitively misleading—miles per
gallon metric). OMB (2013) has issued guidance directing agencies to proactively
experiment with interventions that increase efficiency and efficacy at marginally low
cost, of which interventions derived from ABS are often prime candidates.

**The White House Social and Behavioral Science Team (SBST)**

I’ve spent the last 14 months helping set up a new team within the federal
government to drive the use of applied behavioral science and, more generally, use
experimental methods to evaluate the efficacy and efficiency of programs. Much of the
work (although not all) has focused on identifying low cost opportunities, both in terms
of the interventions and the methods of evaluation. The three examples I give below,
for instance, all involve minor changes to an existing program component (viz., a
website or a letter), and then use already existing administrative data to track the
outcome of interest. In this way, we can learn how to improve the program component at low marginal cost. Here are the three examples:

**Example 1: Industrial funding fee**

**Objective and background**

How can forms be designed to reduce financial self-reporting errors? This is a question of general importance. Here we study a very simple intervention—a signature promising accuracy—within the context of the industrial funding fee, or “IFF.” Federal contractors are required to pay the IFF, currently set at 0.75% of quarterly sales, on all “Multiple Award Schedule” (MAS) transactions—basically, it’s an overhead charge that contractors pay to support the maintenance of the system through which they make their sales. Importantly, the owed IFF is determined almost exclusively via self-reports submitted online (known as the 72a Reporting System); there is virtually no independent auditing of self-reported accuracy. In FY 2013, contractors self-reported about $34.8 billion in MAS sales, of which the General Services Administration (GSA) collected (across roughly 47,000 transactions) approximately $269 million in IFF.

**Research insight**

Experiments have shown that inserting a confirmation prompt, wherein the user signs his or her name confirming the accuracy of self-reported statements, reduces self-report errors if done at the beginning of a form; prompts at the end of a form—as is usually done—seem to have no effect (Shu, Mazar, Gino, Ariely, & Bazerman, 2012). In the experiments by Shu et al., students completed a 20-question math puzzle booklet, for which they earned $1 per correct answer minus a 20% tax. A form (designed
similarly to an IRS Form 1040) elicited self-reported earnings for purposes of exacting the tax. An imperceptible code linked each form to its referenced booklet, so that the researchers could measure cheating. Each student was randomly assigned to receive one of three form versions, which were identical except with regards to a signature line confirming, “I declare that I carefully examined this return and that to the best of my knowledge and belief it is correct and complete”; specifically, that line was either omitted, presented at the bottom of the form, or presented at the top of the form. Students were significantly more likely to cheat if the signed confirmation was either omitted or at the bottom (about 70% cheated) than if presented at the top (only about 40% cheated). Similar results held for a self-report of owed travel expenses.

An additional experiment by Shu et al. (2012) replicated this finding in a field setting. In partnership with an automobile insurance company in the southeast, customers were randomly assigned to receive one of two policy review forms, both of which stated, “I promise that the information I am providing is true,” but varied only as to whether that signature confirmation was on the top or bottom of the form. (Reporting lower odometer mileage indicates less driving, lower risk of accident occurrence, and therefore lower insurance premiums). Across 13,488 completed policy forms for 20,741 cars, customers who signed at the beginning reported higher mileage (about 26,000 miles) than those who signed the bottom (about 23,700 miles)—a significant difference of about 2,300 miles, which the researchers estimated translates into at least a $48 lower annual premium for those who cheated.
Methodology

A randomized controlled trial was fielded during the third reporting quarter of 2014, wherein all contractors in the system (\( N = 18,477 \)) were randomly assigned to use either: (a) the existing 72a Reporting System (control); or (b) a modified 72a interface (treatment), redesigned to include an opening signature box confirming, “I promise that the information I am providing is true and accurate.” Administrative data on paid IFF provided the primary outcome measure.

Results

Amongst contractors remitting a positive fee (\( N = 7,792 \)), the median IFF payments were $536 in the control group and $567 in the treatment groups—a $31 absolute increase, \( p < .05 \), or about a 5% relative increase. At this rate of difference, if all 7,792 vendors used the treatment system rather than the control system, the government would collect about an extra $241,552 per reporting quarter. Recall that the financial cost of this intervention was nothing more than the time taken to add a couple lines of HTML code into the 72a Reporting System website.

Next steps

The evaluation is running again this quarter. This will let us examine whether minor tweaks such as a signature box continue to exert an effect, or if—as some have worried—they might exert a one-time “shock” effect, in the sense that its real power is just to have an unexpected feature prompt extra attention. We might also try, for half of the subjects, moving the signature down to the bottom, simply to demonstrate concretely the importance of where the prompt is located within the process.
Example 2: Influenza vaccine outreach

Objective and background

Influenza (the “flu”) results in more than 120,000 hospitalizations, 36,000 deaths, and $100 billion in direct and indirect spending. Although everyone can be affected by the flu, young children, pregnant women, and the elderly are often at the highest risk of serious and fatal flu complications. The flu vaccine, which is intended for nearly all individuals 6 months of age and older, reduces the likelihood of infection by more than 60%.

Yet only 4 in 10 Americans actually undergo immunization. Studies have identified several reasons. Some barriers are related to information. For example, individuals may believe that if they are typically healthy and rarely fall ill, they are at a decreased risk of infection and therefore do not require an annual vaccination; or they may be unsure about where to get vaccinated or how much it costs. A different category of barrier is motivational. Individuals may believe it is worthwhile to get vaccinated, and they might even have a brief moment of intending to get vaccinated; yet, ultimately, they fail to transform that generalized intention into a specific behavioral plan that is executed. With this in mind, how might behavioral insights be used to help increase flu vaccine update?

Research insight

A set of early studies suggests that simple, low-cost behavioral interventions can help overcome these behavioral barriers to vaccination. Prior research has also shown, in particular, that elaboration of a concrete plan increases the likelihood a person will
execute an intended action. Milkman, Beshears, Choi, Laibson, and Madrian (2011), for example, tested the effect of an “implementation intention prompt,” wherein a letter recipient is invited to write in the day of week, month, day, and time at which he or she plans to get vaccinated. They found that employees at a utility firm were significantly more likely to get vaccinated if receiving an informational mailer with an intention prompt than without one (about 37% versus 33% vaccination rate, respectively).

A different line of work has tested an intervention dubbed “enhanced choice” (Keller, Harlam, Loewenstein, & Volpp, 2011). Recipients are prompted to “decide now”—by checking one of two boxes—indicating they will or will not get vaccinated. The choice is “enhanced” in that the options use language that highlights the losses incumbent in refusing to vaccinate. In Keller et al.’s study, while only 42% of subjects intended to vaccinate when invited to opt-in, 62% and 75% intended to vaccinate when presented an active choice or enhanced active choice, respectively.

Methodology

We randomly assigned Medicare beneficiaries \( N = 200,000 \) to receive one of four different letters aimed at encouraging flu vaccine update. The top and body of the letter was the same across all versions, providing basic information about the flu and the flu vaccine.\(^\text{67}\) What differed was the bottom section, below the signature line,

---

\(^\text{67}\) This language common to all versions embeds several behavioral insights. For example, the risk of death and hospitalization language leverages risk aversion to further motivate action, and the personalization touches (e.g., the signature and picture) aim to capture attention and engagement with the letter. Several of these dynamics could be independently worthy of testing, but we have prioritized different tactics.
where a P.S. tag could be appended. The different letter versions were manipulated as follows:

**Info Only:** For one of the four letters, there was no P.S. addition—the letter simply ended after the signature line. This is referred to as the “Info Only” letter. For the other letters, however, one of three different P.S. tags was appended... 

**Intention Prompt:** An “implementation intention prompt” was provided, inviting the recipient to write in the day of week, month, day, and time at which he or she plans to get vaccinated. Again, prior research has shown that such elaboration of a concrete plan increases the likelihood a letter recipient will get vaccinated (Milkman et al., 2011).

**Enhanced Choice + Intention:** Recipients were prompted to “decide now”—by checking one of two boxes—indicating they will or will not get vaccinated. The choice is “enhanced” in that the options use language that highlights the losses incumbent in refusing to vaccinate (Keller et al., 2011). We also included the intention prompt as part of the affirmative choice.

As for the fourth, a slight narrative detour: In an example of how my new job is somewhat surreal at times, the office of the United States Surgeon General reached out to me while the trial was being designed, explained he was excited about the project, and asked if he could help somehow. We decided to test whether who was the sender had an effect, in this case, either a letter from the Surgeon General or from the Director of the National Vaccine Program Office (NVPO).

Interestingly, people had different intuitions—sometimes strongly conflicting intuitions—about who would be more effective. Some touted the authority and name-brand recognition of the Surgeon General; others worried the government-military overtone would backfire, and the NVPO Director, whose picture just looked like a neighborhood doctor, would be more encouraging. My own take was the picture per se
would be about the same, since it’s likely just an attention effect. People are drawn to
human faces, and having a picture in a letter uses that attention effect to draw the
reader into the letter contents. Who is the face is irrelevant. But what might make a
difference is the letterhead. Medicare beneficiaries receive letters from CMS (Centers
for Medicare and Medicaid) all the time, and might thus be more likely to ignore it as
just a regular humdrum letter (note the NVPO return address would be CMS). People
do not, however, very often get a letter from the Surgeon General, so that might pique
interest more and result in a greater likelihood of opening the letter in the first place.

Regardless, we need not speculate and grasp for theories here. Just test it. And
that’s what we did. The three variants already described (info only; intention prompt;
and enhanced choice + intention) were all sent from the Surgeon General. The fourth
variant was the Info Only letter, but this time with the NVPO Director as the sender. The
next page shows an example of the enhanced choice + intention letter from the Surgeon
General.
Dear [BENEFICIARY.First_Name],

Protect yourself and those you love – get your free flu shot!

Almost 36,000 Americans die every year because of the influenza virus. The “flu” – as it’s commonly called – also results in more than 200,000 hospitalizations annually. Adults 65 and over, children 5 and younger, and pregnant women have the highest risk.

You can help! Getting the flu shot is the best way to protect yourself, your neighbors, and the ones you love. Because flu viruses change from year to year, it’s important to get a flu shot each year.

People with Medicare can get a free flu shot from doctors or other health care providers, and it’s widely available – just visit your local pharmacy, senior center, hospital, or doctor’s office.

So please don’t forget – get your free flu shot today!

Sincerely,

From Boris D. Lushniak

Boris D. Lushniak, M.D., M.P.H.
RADM, U.S. Public Health Service
Acting U.S. Surgeon General

P.S. Many people find it helpful to decide now on a plan for getting their flu shot. Mark your decided plan below, and stick it on your refrigerator so you don’t forget!

☐ I will get the flu shot to reduce my risk of getting and spreading the flu on:

☐ I will not get the flu shot, even if it means I’m more likely to get sick and spread the flu.
Results

This experiment is currently out in the field. Medicare reimbursement claims data will ultimately be used to compare vaccine uptake between the four experimental groups. The key hypotheses are that the intention prompt and enhance choice + intention letters will prompt more persons to get vaccinated. Note, again, how this trial is being run at low marginal cost. Creating the different letter variants used up my personal time, but from the mail center’s perspective was virtually zero cost. And when it comes time to analyze results, we’ll use already existing administrative data from CMS.

Example 3: Curbing CMS fraud, waste, and abuse

Objective and background

The HHS Center for Program Integrity (CPI) is charged with eradicating fraud, waste, and abuse in Medicare. It has recently been undertaking a campaign to identify providers with irregular billing patterns that are suggestive of fraud, waste, or abuse. So identified, they might be subjected to audits or—and here is where we come in—sent information that encourages the provider to re-evaluate their prescribing practice to ensure it is valid and, if necessary, seek help on how to curb wasteful activity. (The thinking here is that much over-prescribing might be accidental; for example, a physician might not realize a practice guideline suggests more conservative treatment). How can those information letters be designed to be most effective?
8. Closing

**Research Insight**

Previous studies have shown that highlighting what is the normal behavior can motivate people to conform to that norm. Simply stating that “9 out of 10 people pay their taxes on time,” for example, has been shown to substantially increase timely tax payments (Hallsworth, List, Metcalfe, & Vlaev, 2014). Effects can be enhanced if a communication explicitly flags the deviance of the recipient, especially relative to peers. Research has demonstrated that home residents use less electricity when informed their energy consumption exceeds that of their neighbors, for instance (e.g., Allcott, 2011).

**Methodology**

Our first experiment is a simple 2×1 design, wherein providers (N = 1,500) were randomized to either receive an informational letter or no letter at all. The first page of the letter was designed, with language and a graphic, to highlight a social norm comparison. We also revamped much of the letter itself, trying to dampen down the typical government legalese and thereby render it a more user-friendly communication. Here is a snapshot of the first page of the letter:
Month XX, 2014

Re: You prescribed **XX% MORE** schedule II controlled substances than your peers.

Dear [NAME],

The figures above display the total count (left) and 30-day equivalent (right) of your Schedule II prescribing, compared to the national and state averages of those within your specialty. As can be seen, you prescribed far more – XX% more – than similar specialists within your state.

We hope that you will use the information provided to see if your high prescribing level is appropriate for your patient population. Read on for more information about the methodology used to analyze your prescribing behavior, and to learn what actions to take next.

Sincerely,

Mark Majestic, Director
Medicare Program Integrity Group
8. Closing

Results

This experiment is currently in the field. We’ll use Medicare claims data to examine physician prescribing patterns. The hypothesis is that those receiving the letter will prescribe less than those who did not receive the letter. We’ll also examine some auxiliary questions, for example if a provider cuts back on their prescribing of schedule II drugs (the focus of the letter), do we see a displacement effect such that they compensate by prescribing more of a different drug or treatment?

Next steps

CPI intends to send out a series of these sorts of communications, so there is much testing opportunity to support learning as to how to best design an effective letter in this context. For example, there is an internal debate about whether these letters are more likely to succeed by embracing a helpful tone (i.e., we’re CMS, and we’re here to help you be a better physician) or an enforcement tone (i.e., we’re CMS, and we’re going to pursue you if you’re breaking the law here). We also, in the first experiment, cannot tease apart a social comparison effect from a pure information effect. A future experiment might allow us to have letters with or without the social comparison.

Applied behavioral science on the horizon

These examples provide a taste of what is on the horizon. The early examples are simple—a small letter tweak, a slight addition to a website—but represent substantial progress in terms of developing a culture of applied behavioral science and experimental testing. For many agency partners, this was the first time they have ever run an experiment. Figuring out how to field a trial in the complex world of the federal
government, especially for the first time, is not a trivial task, and involves a steep learning curve—plus much teaching and convincing that testing is worth the effort and risk. It took about eight months, for example, to work out the industrial funding fee (IFF) experiment. But the upfront investment can have substantial returns. The IFF financial savings were meaningful in themselves, of course; but it also empowered the IFF staff to start thinking differently about their program operations. During one of our conversations about the results, they began to brainstorm other design features they were unsure about and thus wanted to test. And, having put some empirical experience under the belt, we’re now more efficient working together. We designed a second experiment in about eight days, rather than eight months.

Given the many insights from the social and behavioral sciences, and the multifarious ways that the federal government touches on our lives, the opportunities for linking the two as part of creating a more effective and efficient government are virtually limitless. This has, of course, been true for a long time, and behavioral scientists have long discussed the implications of their work. But what is unique is that we find ourselves today in a political environment where many government actors are, for the first time, open to conversation in a serious way, in large part because the public demands and expects it. The President’s management agenda, for instance, has repeatedly stressed the need to ramp up the use of evidence and evaluation across government. This is a tremendous opportunity. It is also a serious responsibility. We pitch certain “nudges”—say, the power of defaults—to illustrate how simple changes can have big impacts. These are important examples. But for most problems,
behavioral insights are one part of a larger puzzle, and we are usually unsure what designs will work best. What we have instead is a body of insights into human psychology and, most importantly, methodological tools for testing how those insights might play out in applied settings. The upshot is that we can’t just toss policy makers clean laboratory findings and hope for the best. We need behavioral scientists getting into the field, getting their hands dirty applying the science in complex policy settings. The creation of the SBST is an example of how social and behavioral scientists are pushing for a seat at the policy table in precisely this way.

References


REFERENCES


California v. Boyette, 58 P. 3d 391 (Supreme Court 2003).

Caperton v. AT Massey Coal Co., Inc., 556 US 868 (Supreme Court 2009).


Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People’s


Guyatt, G. H., Sackett, D. L., Cook, D. J., Guyatt, G., Bass, E., Brill-Edwards, P., ... others. (1994). Users’ guide to the medical literature: II. How to use an article about therapy or prevention. What were the results and will they help me in caring for my patients? *Jama, 271*(1), 59–63.


Magna Trust Co. v. Illinois Cent. R. Co., 728 NE 2d 797 (Appellate Court, 5th Court 2000).


doi:10.1016/j.cognition.2007.07.017


New Jersey Eyewitness Instruction, New Jersey Model Criminal Jury Charges (2012).

Retrieved from

http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf


doi:10.1038/506150a


Patton v. Yount, 467 US 1025 (Supreme Court 1984).


Reynolds v. United States, 98 US 145 (Supreme Court 1879).


Sawyer v. Southwest Airlines Co. (10th Cir. 2005).


Skilling v. US, 561 US 358 (Supreme Court 2010).


State v. Addison, 160 NH 493 (Supreme Court 2010).


doi:10.1038/485298a