EDUCATING JUDGES AND LAWYERS IN BEHAVIORAL RESEARCH: A CASE STUDY

Michael D. Cicchini*
Lawrence T. White**

> "For the rational study of the law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics."[1]
>
> —Oliver Wendell Holmes, Jr.

ABSTRACT

The behavioral sciences play a significant role in shaping the law. Yet, despite their importance, many judges and lawyers harbor serious misconceptions about behavioral research. This Article uses a "case study"—a motion hearing in a criminal case—to educate judges and lawyers in several important behavioral-research concepts.

At the motion hearing, the defense lawyer asked the trial judge to modify the pattern jury instruction on the state's burden of proof. In support of his motion, he cited two behavioral studies. The studies demonstrated that the objectionable part of the pattern instruction—its closing mandate to jurors "not to search for doubt" but "to search for the truth"—lowered the burden of proof below the

* J.D., *summa cum laude*, Marquette University Law School (1999); C.P.A., University of Illinois Board of Examiners (1997); M.B.A., Marquette University Graduate School (1994); B.S., University of Wisconsin - Parkside (1990). Michael Cicchini is a criminal defense lawyer and author.

** Ph.D., University of California, Santa Cruz (1984); M.A., *with distinction*, California State University at Fresno (1979); B.A., *with honors*, Whittier College (1975). Dr. White is Professor of Psychology and Director of the Law & Justice Program at Beloit College.

1. Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 469 (1897).

reasonable-doubt standard. Nonetheless, based on several misconceptions about behavioral research, the judge denied the defense lawyer's motion.

This Article briefly describes the two behavioral studies cited by the defense lawyer. It then identifies and debunks each of the judge's criticisms of the studies and discusses confirmation bias as a potential contributor to the judge's erroneous reasoning. The Article then discusses the role of behavioral sciences in legal decision-making and argues that, when change is warranted, judges should not cling to the status quo simply out of "blind imitation of the past." It then presents a standard framework for assessing the validity of behavioral research, and applies this framework to the two studies that the judge rejected.

TABLE OF CONTENTS

INTRODUCTION

Behavioral research has played a significant role in shaping the law. From Louis Brandeis's brief in *Muller v. Oregon*[2] to Mamie and Kenneth Clark's doll studies in *Brown v. Board of Education*[3] to Gary Gates's studies of same-sex couples in *Obergefell v. Hodges*,[4] appellate courts have frequently relied on behavioral research to guide their decision-making and provide empirical support for their rulings.

Similarly, with regard to criminal law, research on the reliability of eyewitness identifications has affected the admissibility of such evidence at trial and, therefore, has influenced how the police conduct lineups and other identification procedures.[5] And research on polygraphs, police-induced false confessions, forensic interviews of child witnesses, racial bias in capital cases, and juror comprehension of jury instructions has also impacted the law in various ways and to varying degrees.[6]

Behavioral research will continue to play an important role in the law. Consequently, judges and lawyers must attain basic proficiency in scientific and statistical reasoning if they wish to evaluate the reliability of such evidence. This Article aims to advance the goal of scientific and statistical literacy by using a criminal-case motion hearing as a "case study."[7]

At the motion hearing, the criminal defense lawyer asked the trial judge to modify the jury instruction on the state's burden of proof.[8] The judge had intended to use a pattern statewide instruction that identifies the burden as

2. Muller v. Oregon, 208 U.S. 412 (1908); Brief for the State of Oregon, Muller v. Oregon, 208 U.S. 412 (1908) (No. 107), 1908 WL 27605 (AKA "The Brandeis Brief").

3. Brown v. Bd. of Educ., 347 U.S. 483, 494-95 n.11 (1954); Michael Beschloss, *How an Experiment With Dolls Helped Lead to School Integration*, N.Y. TIMES: THEUPSHOT (May 6, 2014), https://www.nytimes.com/2014/05/07/upshot/how-an-experiment-with-dolls-helped-lead-to-school-integration.html (discussing the significance of the doll studies in the context of the Supreme Court's decision to overturn "separate but equal").

4. Obergefell v. Hodges, 135 S. Ct. 2584, 2600 (2015); Brief for Gary J. Gates as Amicus Curiae in Support of Petitioners, Obergefell v. Hodges, 135 S. Ct. 2584 (2015) (Nos. 14–556 et al.), available at https://www.supremecourt.gov/ObergefellHodges/AmicusBriefs/14-556_Gary_ %20J_Gates.pdf.

5. *See generally* ELIZABETH F. LOFTUS ET AL., EYEWITNESS TESTIMONY: CIVIL AND CRIMINAL (5th ed. 2013); *see also* Beth Schuster, *Police Lineups: Making Eyewitness Identification More Reliable*, 258 NAT'L INST. OF JUST. J. 2, 2–8 (2007).

6. *See generally* EDIE GREENE & KIRK HEILBRUN, WRIGHTMAN'S PSYCHOLOGY AND THE LEGAL SYSTEM (8th ed. 2014).

7. *See generally* Transcript of Motion Hearing, State v. Soppa, No. 2016–CM–000940 (Wis. Cir. Ct. dismissed Oct. 16, 2017) [hereinafter Hearing Transcript] (on file with the authors).

8. *Id*. at 13.

"beyond a reasonable doubt," but then concludes by telling the jury: "you are *not* to search for doubt. You are to search for the truth."[9]

The defense lawyer asked the judge to remove this closing mandate from the jury instruction. Telling the jury not to search for doubt is impermissible, as it is the jury's *duty* to "examin[e] the evidence for reasonable doubt."[10] Further, "'seeking the truth' suggests determining whose version of events is more likely true, the government's or the defendant's, and thereby intimates a preponderance of evidence standard."[11] This, of course, would appear to violate due process, which "protects the accused against conviction except upon proof beyond a reasonable doubt."[12]

When the defense lawyer asked the judge to delete the pattern instruction's closing mandate, he presented the legal arguments outlined above. In support of this request, the lawyer cited to our two behavioral studies that were recently published in the *University of Richmond Law Review*[13] and the *Columbia Law Review Online*.[14] Both studies directly tested the impact of the contentious portion of the instruction—the directive "not to search for doubt" but "to search for the truth"—on mock juror verdicts. As discussed in the next Part, both studies provided empirical evidence that the closing mandate does, in fact, lower the government's burden of proof.

Despite the defense lawyer's arguments and his citation to our published studies,[15] the judge denied the motion. The judge based his ruling on perceived inadequacies of the studies.[16] When ruling, the judge claimed that he understood the social sciences and statistics involved. However, as will be demonstrated, we believe that the judge misunderstood these principles of behavioral research.

This Article proceeds as follows. Part I briefly describes our published studies that the defense lawyer cited in support of his motion. Part II, the heart of the Article, seeks to debunk the judge's seven criticisms of the studies. Part III then offers an explanation for the judge's misguided reasoning, and discusses

9. WIS. CRIM. JURY INSTRUCTION No. 140 (2016) (emphasis added).

10. State v. Berube, 286 P.3d 402, 411 (Wash. Ct. App. 2012) (stating that "[i]n a criminal case, the State must prove its case beyond a reasonable doubt. The jury cannot discern whether that has occurred without examining the evidence for reasonable doubt.").

11. United States v. Gonzalez–Balderas, 11 F.3d 1218, 1223 (5th Cir. 1994).

12. *In re* Winship, 397 U.S. 358, 364 (1970).

13. Michael D. Cicchini & Lawrence T. White, *Truth or Doubt? An Empirical Test of Criminal Jury Instructions*, 50 U. RICH. L. REV. 1139 (2016) [hereinafter *Empirical Test*].

14. Michael D. Cicchini & Lawrence T. White, *Testing the Impact of Criminal Jury Instructions on Verdicts: A Conceptual Replication*, 117 COLUM. L. REV. ONLINE 22 (2017) [hereinafter *Conceptual Replication*].

15. The defense lawyer who filed the motion is not affiliated with defense lawyer Michael D. Cicchini, who co-authored the two behavioral studies and this Article.

16. *See generally* Hearing Transcript, *supra* note 7, at 13–18.

how similar misconceptions can be avoided in the future. Finally, Part IV presents a standard framework used to assess the validity of behavioral studies; it then uses this framework to evaluate the two studies that the judge rejected.

I. A TALE OF TWO STUDIES

In the first behavioral study cited by the defense lawyer, 298 participants served as mock jurors in a hypothetical criminal case.[17] Each juror read a criminal-case summary that included the elements of the charged crime, a summary of the witnesses' testimony, and the lawyers' closing arguments.[18] Prior to rendering verdicts, jurors were randomly assigned to three groups, and each group was provided with a different instruction on the burden of proof.[19]

Jurors in the first group were instructed to search for the truth, with no mention of the reasonable-doubt standard; 29.6% of these jurors voted to convict.[20] Jurors in the second group received a legally proper reasonable-doubt instruction, with no mention of searching for the truth; only 16% of these jurors voted to convict.[21] Jurors in the third group—the crucial test group—were properly instructed on reasonable doubt and then told "not to search for doubt" but "to search for the truth."[22] 29% of these jurors voted to convict—a rate statistically indistinguishable from that of jurors in the first group, who received no reasonable-doubt instruction whatsoever.[23]

The second study cited by the defense lawyer was a conceptual replication of the first.[24] In this study, 248 mock jurors read a criminal-case summary and were randomly assigned to one of two groups.[25] Jurors in the first group received a standard reasonable doubt instruction; 22.6% voted to convict.[26] Jurors in the second group received the same reasonable-doubt instruction, plus the mandate "not to search for doubt" but "to search for the truth"; 33.1% voted to convict.[27]

After rendering a verdict, mock jurors in the second study were asked if conviction would be proper even if they had a reasonable doubt about the

17. *Empirical Test*, *supra* note 13, at 1151.
18. *Id*.
19. *Id*. at 1152.
20. *Id*. at 1152, 1154.
21. *Empirical Test*, *supra* note 13, at 1152, 1154.
22. *Id*. at 1153
23. *Id*. at 1155.
24. *Conceptual Replication*, *supra* note 14, at 27.
25. *Id*. at 28–29.
26. *Id*. at 29–31.
27. *Id*.

defendant's guilt. [28] Compared to the standard reasonable-doubt group, jurors in the search-for-truth group were nearly twice as likely to hold this mistaken belief (28% and 15% respectively).[29] Further, jurors who held this mistaken belief, regardless of the jury instruction they received, voted to convict the defendant at a significantly higher rate than jurors who properly understood the burden of proof (54% and 21% respectively).[30]

In summary, two published studies were cited and provided empirical support for the defense lawyer's argument that telling jurors "not to search for doubt" but "to search for the truth" lowers the burden of proof below the constitutionally-mandated standard.[31] To our knowledge, these two studies are the only studies to test the impact of this particular language on jurors' verdicts.

II. CRITIC'S CORNER: THE JUDGE WEIGHS IN

Based on his own critique of the two studies, the judge denied the defense lawyer's motion to modify the jury instruction.[32] As we demonstrate below, the judge's analysis was misguided, yet ultimately instructive. It vividly illustrates several misconceptions about behavioral research. We are unaware of how pervasive these misconceptions are in the legal field. However, studies indicate that "judges are not especially skilled at distinguishing between high-quality and low-quality research,"[33] and the trial judge's remarks during the motion hearing implied that other judges share his opinions.[34]

28. *Conceptual Replication*, *supra* note 14, at 31–32.

29. *Id*. at 32.

30. *Id*.

31. *See Empirical Test*, *supra* note 13, at 1166; *Conceptual Replication*, *supra* note 14, at 35.

32. *Hearing Transcript*, *supra* note 7, at 20.

33. MARK COSTANZO & DANIEL KRAUSS, FORENSIC AND LEGAL PSYCHOLOGY: PSYCHOLOGICAL SCIENCE APPLIED TO LAW 15 (2d ed. 2015).

34. During the hearing, the judge recalled a recent discussion he and another judge had regarding the subject, noting:

> We also talked about [the jury instruction] at the seminar I was at last week in the [Wisconsin] Dells for the judges. The Wisconsin Criminal Jury Instruction Committee addressed us, and they addressed this very issue, and . . . they may not have gone into the statistics . . . but I was happily sitting next to a judge who's on the Court of Appeals who has a statistical background as well, and he and I talked about all this stuff, and he and I both agreed on it.

Hearing Transcript, *supra* note 7, at 17.

A. *"A Dog in the Fight"*

When denying the defense lawyer's motion, the judge rejected the studies cited in support of the motion because one of the researchers is a defense attorney. In particular, the judge stated:

> [U]ntil other researchers have—who don't have a dog in the fight, as it were, who are *not defense attorneys*, who are neutral, for example, somebody from the University of Cincinnati that does seem to do a lot of this material looked at that and studied that one issue, then I might consider it.[35]

When the judge dismissed the studies based on the occupation of one of the researchers, he committed a logical error known as the *ad hominem* fallacy.

> The *ad hominem* fallacy involves bringing negative aspects of [a researcher], or their situation, to bear on the view they are advancing. . . . [A] subtle version of the fallacy is the circumstantial *ad hominem* in which, given the circumstances in which the [researcher] finds him or herself, it is alleged that their position is supported by self-interest rather than by good evidence. Hence, the scientific studies produced by industrialists to show that the levels of pollution at their factories are within the law may be undeservedly rejected because they are thought to be self-serving. Yet it is possible that the studies are sound: just because what someone says is in their self-interest does not mean it should be rejected.[36]

As the information above suggests, it is common to find a person doing research on a subject in which he or she has an interest. In fact, the scenario demanded by the judge—where a researcher has no interest in the subject under investigation—is likely *uncommon*. In real life, drug companies test vaccines; environmentalists study climate change; politically-oriented think tanks (both left- and right-leaning) study tax policy and recommend reforms. A study may be valid or invalid, but its validity does not depend on the researcher's employment.

---

35. *Id.* at 16 (emphasis added). It is important to note, the judge did not realize the second author of the studies at issue is not a lawyer. On the contrary, he is a researcher and college professor holding a Ph.D. in Psychology—albeit not from the University of Cincinnati.

36. Hans Hansen, *Fallacies*, STAN. ENCYCLOPEDIA OF PHIL. (May 29, 2015), http://plato.stanford.edu/entries/fallacies/; *see also* Bradley Dowden, *Fallacies*, INTERNET ENCYCLOPEDIA OF PHIL., http://www.iep.utm.edu/fallacy/_ (last visited Oct. 18, 2017) (equating the circumstantial *ad hominem* with "Guilt by Association").

Equally important, defense lawyers are not the only legal actors who have an interest in ensuring that the jury instruction accurately conveys the constitutionally-mandated burden of proof. First, every prosecutor is a "minister of justice," a role that "carries with it specific obligations to see that the defendant is accorded procedural justice . . . ."[37] Second, even judges have "a dog in the fight," as a court is obligated to "exercise its discretion in order 'to fully and fairly inform the jury of the rules of law applicable to the case and to assist the jury in making a reasonable analysis of the evidence.'"[38]

To exclude defense lawyers, prosecutors, and judges from a debate about jury instructions, simply because they have an interest in the subject, would lead to an untenable and somewhat absurd state of affairs. Consequently, the judge's *ad hominem* attack on the researcher, instead of the research, misses the mark.

B. *The Quality of the Participants*

When denying the defense lawyer's motion and rejecting the two studies, the judge criticized the studies' participants: "We don't know the nature of the quality of the people he's using. . . . They may all be essentially college students with a certain amount of education that may skew the quality of . . . what is being produced."[39]

On its face, the judge's concern about the study participants being non-representative is a legitimate one. Many studies, especially those conducted by research psychologists at universities, draw their participants from a readily available pool of college students.[40] This common study-design feature has

37. MODEL RULES OF PROF'L CONDUCT r. 3.8 cmt. 1 (AM. BAR. ASS'N 2017). A vast majority of states have adopted rules of professional conduct that, while varied in substance, closely mirror the ABA's Model Rules of Professional Conduct. *See* State Adoption of the ABA Model Rules of Professional Conduct, AM. BAR. ASS'N, https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/alpha_list_state_adopting_model_rules.html (last visited Jan. 11, 2018) (listing the jurisdictions that have adopted the model rules).

38. State v. Neumann, 832 N.W.2d 560, 584 (Wis. 2013) (citation omitted); *see also* State v. Redmond, 78 P.3d 1001, 1003 (Wash. 2003) ("Parties are entitled to instructions that, when taken as a whole, properly instruct the jury on the applicable law, are not misleading, and allow each party the opportunity to argue their theory of the case.") (citation omitted); People v. Hernandez, 107 Cal. Rptr. 3d 915, 919 (2010) ("The trial court has a sua sponte duty to instruct the jury on the general principles of law relevant to the issues raised by the evidence. . . . Additionally, even if the court has no sua sponte duty to instruct on a particular legal point, when it does choose to instruct, it must do so correctly.") (citations omitted).

39. *Hearing Transcript*, *supra* note 7, at 15.

40. KEITH E. STANOVICH, HOW TO THINK STRAIGHT ABOUT PSYCHOLOGY 117 (10th ed. 2013) (stating "college sophomores are the subjects in an extremely large number of psychological investigations . . . .").

prompted discussion of the "college sophomore problem."[41] That is, if controlled studies use exclusively college students, we cannot know if the findings will generalize to other demographics.

The judge remarked that he had thoroughly read the two studies. "Believe me," he said, "I researched the heck out of them."[42] However, the judge seemingly never addressed that both studies described the test participants in detail. In the first study, 52% of the participants were male, with an average age of 37 years.[43]Additionally, 19% of the participants were non-White, and 14% had prior jury experience.[44] They came from forty-two different states, including the District of Columbia, and 40% had already graduated from a four-year college.[45]

In the second study, 48% of the participants were male, with an average age of 35.8 years.[46] Moreover, 26% were non-White, and 13% had prior jury experience.[47] Further, they came from forty-two different states, and 56% had already graduated from a four-year college.[48] In short, the judge's criticism about a "college sophomore problem" was misplaced.

Nonetheless, even assuming the judge was correct that the mock jurors in the two studies were all college students, the studies' main conclusion—that telling jurors "not to search for doubt" but "to search for the truth" lowers the burden of proof—remains valid for three reasons:

First, "[t]he college sophomore criticism does not *invalidate* past results, but simply calls for *more* findings that will allow assessment of the theory's generality."[49] If college students respond oddly to the test variable, the problem will reveal itself when subsequent studies employ participants who are not college students.

Second, with regard to jury instruction experiments in particular, there is little evidence that jurors' personal characteristics matter. In a comprehensive review of 206 jury decision-making studies, the authors concluded that "[i]t has been a source of discouragement to some, and relief to others, that so many

41. *Id*. at 117–21 (discussing the college sophomore problem in psychological findings).

42. *Hearing Transcript*, *supra* note 7, at 19.

43. *Empirical Test*, *supra* note 13, at 1151.

44. *Id*.

45. *Id*.

46. *Conceptual Replication*, *supra* note 14, at 28.

47. *Id*.

48. *Id*.

49. STANOVICH, *supra* note 40, at 117 (emphasis original) (this is one of several "legitimate responses" to the college sophomore criticism).

studies have yielded so little evidence that individual verdict preferences are reliably predicted by personal characteristics."[50]

Third, with *experiments* of any kind, as opposed to *surveys*, the college-sophomore problem is not particularly an issue.[51] Specifically, surveys use a subset of a population to forecast the frequency of a specific characteristic—for example, support for a political candidate—in the larger population.[52] To be useful, surveys must possess a high degree of external validity—a concept discussed in Part IV. In other words, the sample used in the survey must be representative of the population. A survey solely of college students may not be sufficient to draw conclusions with a high degree of generalizability to the larger population.

Experiments, on the other hand, do not use the test participants to forecast the frequency of a specific characteristic in the larger population. Rather, experiments are interested in the differences between the test groups.[53] For example, in our jury instruction experiments, we had no interest in forecasting the frequency of guilty verdicts in real-life cases. The reasoning for this is because overall conviction rates will vary wildly based on the unique facts of any given case. Rather, we were interested in learning the differences between test groups—for example, all else being equal, did mock jurors who received instruction A vote guilty more often than jurors who received instruction B?

Therefore, because the participants in the two studies *were* very diverse, and because the studies were experiments rather than surveys, the judge's criticism about "the nature of the quality of the people" used in the studies was misplaced.

C. *"Lies, Damn Lies, and Statistics"*

When denying the defense lawyer's motion, the judge also criticized the two studies for reporting statistics in a deceptive manner. The judge stated:

> [T]here's some very misleading things in the report. For example, I think it's—and I believe it was Mark Twain said [sic], "Lies"—no. "Lies, damn lies, and statistics" is what I think he said. To say that one group

50. Dennis J. Devine et al., *Jury Decision Making – 45 Years of Empirical Research on Deliberating Groups*, 7 PSYCHOL., PUB. POL'Y, & L. 622, 700 (2000).

51. *See* Beth Morling, RESEARCH METHODS IN PSYCHOLOGY: EVALUATING A WORLD OF INFORMATION 173 (2012) (discussing how sample selection is far more important for a survey, or "frequency claim," than it is for controlled experiments that seek to detect "associations and causes").

52. JANET M. RUANE, INTRODUCING SOCIAL RESEARCH METHODS: ESSENTIALS FOR GETTING THE EDGE 232–33 (2016).

53. PAULA BERINSTEIN, FINDING STATISTICS ONLINE: HOW TO LOCATE THE ELUSIVE NUMBERS YOU NEED 20 (1998).

> had a 50 percent increase of probability of guilty simply because in one group it was 20 that voted guilty and 30 in the other is misleading by a large extent.[54]

The phrase "lies, damn lies, and statistics" [55] is often invoked when one does not like a conclusion that is supported by statistical data.[56] In our example, the judge was not fond of the statistical findings of the cited studies, so he challenged their credibility by associating the statistics with lies. However, attacking the integrity of statistics because they conflict with one's beliefs on an issue is the equivalent of condemning the English language because one does not agree with the point of a sentence.

The judge's claim of "misleading things in the report" was inaccurate in two respects. First, contrary to the judge's assertion, the studies did not estimate the probability that a certain percentage of jurors would vote guilty in future cases. Rather, the studies simply reported how many mock jurors voted guilty in each of the test conditions.[57] The distinction between findings (what has happened) and predictions (what will happen) is an important one, as noted in the previous Part.

Second, when the judge referred to 20 guilty votes in one group and 30 guilty votes in another group, and said it was "very misleading" to describe the difference as a "50 percent increase," he was likely referring to the second study, which reported the following:

> In Group 1, which received the doubt-only instruction, only 28 of 124 mock jurors returned verdicts of guilt for a group conviction rate of 22.6%. In Group 2, which received the doubt-and-truth instruction, 41 of 124 mock jurors returned verdicts of guilt for a group conviction rate of 33.1%. That is, the conviction rate among jurors who were told "not to search for doubt" but instead "to search for the truth" was *almost 50% higher* than the conviction rate for jurors who were simply instructed to evaluate the state's case for reasonable doubt.[58]

54. *Hearing Transcript*, *supra* note 7, at 15–16.

55. In fact, the quote is frequently misattributed to Mark Twain. *See* Paul F. Velleman, *Truth, Damn Truth, and Statistics*, 16 J. STAT. EDUC. 1, 2 (2008) (Although he may have said it, "Twain did not originate it (nor, of course, did he claim to), and he was most likely mistaken in attributing it to Disraeli.").

56. *See id*. at 1 (noting a Google search for "lies, damn lies, and statistics" returned "about 207,000" results, with most designed to "suggest dishonest manipulations and interpretations.").

57. *Empirical Test*, *supra* note 13, at 1154–55; *Conceptual Replication*, *supra* note 14, at 29–31.

58. *Conceptual Replication*, *supra* note 14, at 30–31 (emphasis added).

The judge did not explicitly state *why* he thought it was "very misleading," or "misleading by a large extent" to describe a change from 20 to 30 as a "50 percent increase." Perhaps the judge was thinking of the classic case of an attention-grabbing headline that states, for example, "Drug X doubles your risk of brain cancer." The claim might technically be true. If the brain cancer rate in the population is .001, but increases to .002 among those who use the drug, the claim that the rate doubled is true, but it is also misleading because the risk of brain cancer is incredibly low in either case.

Conversely, as illustrated above, the study the judge labeled as "misleading" provided the number of convictions in each group, the size of each group, the conviction rate of each group, and the percentage increase for the group that was instructed "not to search for doubt" but "to search for the truth." Further, unlike the brain cancer example, an increase in conviction rates from 22.6% to 33.1% is both meaningful in a practical sense and, as explained in the next Part, statistically significant. The increased conviction rate, due to a flawed jury instruction, may have been an inconvenient concept for the judge. However, the statistical information in the studies supported this assertion.

D. *Sample Size and Significance*

When denying the defense lawyer's motion, the judge opined, "It's a 250-person study. It's still not significant . . . . From a statistical point of view, and I've had a lot of statistics in college, this is a not significant amount of people. It's not big enough, it's too small of a study."[59] This statement raises two questions. First, what were the sample sizes of each of the studies the judge was criticizing? Second, were the samples too small?

The first study had 298 mock jurors divided into three groups.[60] The second study had 248 mock jurors divided into two groups.[61] The number of participants *per condition* ranged from 98 to 124.[62] To put these numbers in perspective, it is instructive to compare them to other published studies that have examined the impact of reasonable-doubt instructions on mock juror decision-making.

In a 1985 report of three experiments, a first experiment divided 198 undergraduates into seven groups (six experimental conditions with 29 participants in each, and a control condition containing of 24); a second experiment divided 220 undergraduates into six groups (36 or 37 per condition); a third experiment divided 96 undergraduates into three groups (32 per

59. *Hearing Transcript*, *supra* note 7, at 13–14.

60. *Empirical Test*, *supra* note 13, at 1152.

61. *Conceptual Replication*, *supra* note 14, at 29.

62. *Empirical Test*, *supra* note 13, at 1154; *Conceptual Replication*, *supra* note 14, at 29.

condition).[63] In a 1996 study, 480 jury-eligible adults were divided into five groups with an average of 96 participants per condition.[64] In a 2007 report of two experiments, the first experiment divided 26 undergraduates into two groups (13 per condition); the second divided 172 undergraduates into two groups (86 per condition).[65] More recently, in a 2015 study, 200 adults were divided into four groups, with an average of 50 participants per condition.[66]

The two studies criticized by the judge as "too small" with "a not significant amount of people" had more participants per condition than any of the other similar studies we were able to locate. (The two studies also used participants who were more diverse in terms of age and education than participants in two of the four studies described above.)[67] Therefore, the judge was incorrect when he said that the samples were "too small" and, more importantly, made an additional mistake: he equated sample size with statistical significance. Statistical significance is partly dependent on sample size, but they are not the same thing.

Statistical significance refers to the degree of confidence that an observed effect—for example, the difference in conviction rates between test groups—actually exists and did not occur by chance.[68] Statistical significance is a function of sample size *and* effect size.[69] Large effects can be reliably detected with relatively small samples.[70]

The statistical significance of an observed effect is expressed by a statistic called the *p*-value.[71] The lower the *p*-value, the more confident a researcher is that his or her findings did not occur by mere chance.[72] If the *p*-value is .05 or

63. Dorothy K. Kagehiro & W. Clark Stanton, *Legal vs. Quantified Definitions of Standards of Proof*, 9 L. & HUM. BEHAV. 159, 163, 165–66, 170 (1985).

64. Irwin A. Horowitz & Laird C. Kirkpatrick, *A Concept in Search of a Definition: The Effects of Reasonable Doubt Instructions on Certainty of Guilt Standards and Jury Verdicts*, 20 L. & HUM. BEHAV. 655, 659, 663 (1996).

65. Daniel B. Wright & Melanie Hall, *How a "Reasonable Doubt" Instruction Affects Decisions of Guilt*, 29 BASIC & APPLIED SOC. PSYCHOL. 91, 93, 95 (2007).

66. Mandeep K. Dhami et al., *Instructions on Reasonable Doubt: Defining the Standard of Proof and the Juror's Task*, 21 PSYCHOL., PUB. POL'Y, & L. 169, 172–173 (2015).

67. *Compare Empirical Test*, *supra* note 13, at 1151 (age ranging from 19 to 76 years, 40% possessing a college degree and 36% having completed some college), *and Conceptual Replication*, *supra* note 14, at 28 (age ranging from 19 to 73 years, 56% possessing a college degree and 35% having completed some college), *with* Kagehiro & Stanton, *supra* note 63, at 162, 165, 170, *and* Wright & Hall, *supra* note 65, at 93, 95 (selecting subjects exclusively from the undergraduate student population at local universities).

68. ARTHUR ARON & ELAINE N. ARON, STATISTICS FOR PSYCHOLOGY 92 (3d ed. 2003).

69. *Id*. at 271.

70. *Id*. at 285.

71. *Id.* at 92; *see also* PAUL C. STERN & LINDA KALOF, EVALUATING SOCIAL SCIENCE RESEARCH 150 (2d ed. 1996).

72. STERN & KALOF, *supra* note 71, at 150.

lower, it is considered statistically significant; *p*-values of .02 or lower are considered highly significant.[73]

In the studies criticized by the judge, the sample sizes were *large* and the effect sizes were *moderate*, resulting in *p*-values that ranged from .033 (significant)[74] to less than .001 (highly significant).[75] In plain language, the differences in conviction rates produced by the different instructions almost certainly did not occur by chance.

For purposes of this Part, the gist of the lesson is clear: Even though the two studies *did* use very large samples, statistical significance cannot be determined on the basis of sample size alone. Rather, statistical significance is indicated by the value of *p*, not the number of participants.

E. *Participant Bias*

In his remarks in open court, the judge expressed doubts about the legitimacy of the studies' findings because the mock jurors may have been biased in some way. He stated "it does not appear as though [the first author] has taken any of the other confidences necessary to assure us that the reason why they went the way they did is based upon the simple words at the end of the 140 instruction."[76] Further on, the judge opined "[y]ou know, for example, do they have an interest in the case? You know, how is their intelligence? Did they essentially look like they had a bias? Well, anything of that sort, and none of that is referred to in this case."[77]

As a matter of fact, several kinds of bias can wreak havoc with behavioral research.[78] Researchers strive to eliminate and control these biases whenever possible. One type of bias is "selection bias," which occurs when participants in a study are recruited in such as a way as to compromise the representativeness of the study sample.[79] A good way to combat selection bias is to randomly select participants from the population to which the researcher wants to generalize the

73. For a general discussion of sample size, effect size, statistical significance, and *p*–values, see ARON & ARON, *supra* note 68.

74. *Conceptual Replication*, *supra* note 14, at 31 (p=0.033 for the "difference in conviction rates between the two groups.").

75. *Id*. at 32 (p<.001 for the difference in conviction rates between jurors with correct versus mistaken understanding of the burden of proof).

76. *Hearing Transcript*, *supra* note 7, at 13–14.

77. *Id*. at 17.

78. *See* ROYCE A. SINGLETON, JR. & BRUCE C. STRAITS, APPROACHES TO SOCIAL RESEARCH 32 (3d ed. 1999).

79. *See* STERN & KALOF, *supra* note 71, at 86 (acknowledging "[a] *biased* sample is one that contains a *systematic error*; it is consistently different from the population in a particular direction.").

study's findings.[80] (As explained in Part II.B., this is a concern for simple surveys but typically is not a concern for experiments—particularly those testing the impact of jury instructions.)

Another type of bias is "participant bias," which occurs when participants in an experiment behave in a way intended to support (or sabotage) the researcher's hypothesis.[81] To combat participant bias, the researcher can make it difficult or impossible for participants to guess the purpose of the study.[82] Related to this is "experimenter bias." This type of bias occurs when a researcher inadvertently acts in a way that influences the participants.[83] To combat experimenter bias, the researcher can give the responsibility for data collection to someone who does not know the hypothesis that is being tested.[84] Such double-blind experiments, in which participants *and* data collectors are unaware of the study's purpose, are highly desirable because they control for *both* participant bias and experimenter bias.[85]

The two behavioral studies criticized by the judge were conducted via Amazon's Mechanical Turk online behavioral research platform.[86] Behavioral scientists, including jury researchers, have used Mechanical Turk to conduct valid online studies of decision-making.[87] Typically, in studies of juror decision-

80. *Id*. (explaining "[t]he only way to be certain that a sample is representative is to use a truly *random sample*.").

81. *See* C. JAMES GOODWIN, RESEARCH IN PSYCHOLOGY: METHODS AND DESIGN 234 (6th ed. 2009); *see also* JOEL ROBERT DAVITZ & LOIS LEIDERMAN DAVITZ, EVALUATING RESEARCH PROPOSALS: A GUIDE FOR THE BEHAVIORAL SCIENCES 17–18 (1996) (regarding participant bias as an "important source of uncontrolled variance" that must be acknowledged in planning research and in the interpretation of the results).

82. GOODWIN, *supra* note 81, at 234 ("The primary strategy for controlling participant bias is to reduce demand characteristics to the minimum. One way of accomplishing this, of course, is through deception.").

83. *See* DAVITZ & DAVITZ, *supra* note 81, at 17–18 (noting "experimenter biases" can "affect a researcher's activities insofar as they influence the questions he or she investigates and the ways in which he or she formulates research problems.").

84. *See* WILLIAM J. RAY, METHODS TOWARD A SCIENCE OF BEHAVIOR AND EXPERIENCE 256–57 (9th ed. 2009).

85. *See* DAVITZ & DAVITZ, *supra* note 81, at 72 (noting the "procedure controls for self–fulfilling prophecies by giving the researcher in contact with the subject the same expectancy for all subjects. It controls for the placebo effect by giving all subjects the same expectations of help."); STERN & KALOF, *supra* note 71, at 72 (explaining a double-blind experiment preserves equal expectations of outcomes for both researchers and subjects).

86. *Hearing Transcript*, *supra* note 7, at 13–15; *Conceptual Replication*, *supra* note 14, at 25, 28; *Empirical Test*, *supra* note 13, at 1150.

87. *See* Winter Mason & Siddharth Suri, *Conducting Behavioral Research on Amazon's Mechanical Turk*, 44 BEHAV. RES. 1, 1–2 (2011) (discussing psychologists' use of Amazon's Mechanical Turk research platform for "easy access to a large, stable, and diverse subject pool . . . ."); *see also* Christina M. O'Donnell & Martin A. Safer, *Jury Instructions and*

making, participants presumably know that some variable is being manipulated (tested). However, it may be nearly impossible for them to divine *which* variable is being tested—the kind of case, the strength of the evidence, the defendant's age or gender or ethnicity, the judge's instructions, and so forth. Participants in online studies do not interact with the researcher or the researcher's assistant; the study materials are presented automatically and in a uniform fashion.[88] In short, on-line studies of juror decision-making are double-blind studies that largely eliminate the possibility of participant bias and experimenter bias.

While the judge did not articulate his bias-related criticisms clearly, he seemed to be concerned that the mock jurors' verdicts may have been influenced by factors other than, or in addition to, the different reasonable-doubt instructions. He was correct in the belief that human decisions and judgments are affected by a host of factors. However, he was incorrect when he described this phenomenon as a kind of "bias" that undermines the validity of controlled experiments. As explained in the first study, this is why participants are randomly assigned to different test groups.

> The virtue of random assignment is that, when used with large numbers of study participants, it produces groups that are statistically equivalent to each other in all respects. Each group has roughly the same number of mock jurors, the same number of men and women, the same number of well–educated and poorly educated persons, and *the same number of biased and unbiased individuals*.
>
> When test groups are statistically equivalent at the outset, receive different jury instructions, and then convict at different rates, we can be quite certain that the different conviction rates were produced by the different jury instructions and not by personal characteristics of the mock jurors in a particular group. In plain language, random assignment creates a level playing field where the *effects of bias are distributed equally across the test conditions*. Therefore, *the end result—a difference in conviction rates—can only be attributed to the type of jury instruction received*.[89]

Mock–Juror Sensitivity to Confession Evidence in a Simulated Criminal Case, 23 PSYCHOL., CRIME & L. 946, 951 (2017) (utilizing the Mechanical Turk research platform).

88. *See* Matthew J.C. Crump et al., *Evaluating Amazon's Mechanical Turk as a Tool for Experimental Research*, 8(3) PLOS ONE e57410 2 (2013), http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0057410&type=printable (stating in regard to online data collection, "the experimenter never directly meets or interacts with the anonymous participants, it minimizes the chance the experimenter can influence the results.").

89. *Empirical Test*, *supra* note 13, at 1165 (emphasis added).

In other words, because we were comparing conviction rates among groups that received different instructions (and not estimating conviction rates in actual jury trials), and because large sample sizes and random assignment yield equal numbers of "biased and unbiased individuals" in each of the groups, the judge's concern about biased participants is misplaced. The fact that some participants in our study may have been biased in some unspecified way cannot explain the different conviction rates observed in the groups.

F. *Juror Deliberations*

In addition to his other stated concerns, the judge also rejected the two behavioral studies because the mock jurors did not deliberate before rendering their verdicts.[90] Instead, participants rendered their verdicts immediately after reading the case-summary materials.[91] The judge explained, "[i]t's simply them going through, reading the materials . . . and hearing what the jury instructions are and then getting their opinion, and it completely does away with the deliberative process that is so important in the jury process . . . ."[92] On its surface, this appears to be a legitimate concern. However, the judge's criticism was misplaced for three reasons.

First, juror decision-making studies *without* deliberations are very common in behavioral research.[93] In one study, for example, undergraduates read their case-summary materials and rendered verdicts without deliberations to test what influence juror gender differences and disabilities in infant victims had on jurors' reactions in infanticide cases.[94] In another, participants rendered verdicts without deliberations to test decision-making when variables such as "victim gender, defendant gender, and defendant age . . . were manipulated."[95]

90. *Hearing Transcript*, *supra* note 7, at 14–15.

91. *Id*.; *Empirical Test*, *supra* note 13, at 1162; *Conceptual Replication*, *supra* note 14, at 28–30.

92. *Hearing Transcript*, *supra* note 7, at 17.

93. RON C. MICHAELIS ET AL., A LITIGATOR'S GUIDE TO DNA: FROM THE LABORATORY TO THE COURTROOM 243 (2008) (stating that a difference between real trials and mock juries is that "in mock jury studies, the jurors usually answer without deliberating with other jurors.").

94. Bette L. Bottoms et al., *Gender Differences in Jurors' Perceptions of Infanticide Involving Disabled and Non-Disabled Infant Victims*, 35 CHILD ABUSE & NEGLECT 127, 128, 132 (2011).

95. Joanna D. Pozzulo et al., *The Effects of Victim Gender, Defendant Gender, and Defendant Age on Juror Decision Making*, 37 CRIM. JUST. & BEHAV. 47, 47, 54 (2010).

Similarly, of the four other studies of reasonable-doubt instructions discussed above,[96] only one discussed using deliberations.[97] In that study, participants were placed into six-person juries that deliberated for up to 90 minutes.[98] The participants' self-reported interpretations of the reasonable-doubt standard were *unaffected* by deliberations.[99]

Second, in wider literature, evidence is mixed about the impact of deliberations. The first of the two studies criticized by the judge addressed this issue:

> [S]everal studies have tested the impact of deliberations on the physical attractiveness bias, examining the tendency for jurors to perceive and treat attractive defendants more favorably than plain-looking defendants. A study in 1974 found that deliberation mitigated the physical attractiveness bias. A study in 1990, however, found that deliberation exacerbated the bias.[100]

Third, and most significantly, the second study criticized by the judge included this finding: mock jurors who were told "not to search for doubt" but "to search for the truth" were nearly *twice* as likely to mistakenly believe that it was proper to convict the defendant even if they had a reasonable doubt about guilt.[101] As discussed earlier, trial judges are duty-bound to clearly and accurately instruct jurors on the government's burden of proof.[102] Overall, it seems odd to choose to instruct jurors in a way that creates a serious, mistaken belief about the burden of proof, only to *hope* that the misconception will be corrected later during jury deliberations.

In summary, published studies without deliberations are very common, the impact of deliberations is at best unclear, and judges should be *very* concerned about the impact of jury instructions on jurors' beliefs and understandings *before* they begin deliberations.

---

96. Discussion *supra* Part II.D.

97. Horowitz & Kirkpatrick, *supra* note 64, at 661 (stating "juries were sent to individual 'jury rooms' to deliberate.").

98. *Id*. at 661, 665.

99. *Id*. at 663.

100. *Empirical Test*, *supra* note 13, at 1163 (internal citations omitted); *see also* MICHAELIS ET AL., *supra* note 93 (stating that "[i]t is hard to know what effect group deliberations will have on an individual's reasoning, especially when opposing fallacies collide. . . . When error rates are high, however, as mock jury research suggests in real trials, group deliberations often foster the exchange of misinformation.").

101. *Conceptual Replication*, *supra* note 14, at 31–32.

102. Discussion *supra* Part II.B.

G. *Peer Review*

The judge also scrutinized the two behavioral studies, and denied the defense lawyer's motion, because the judge thought the studies had not been peer reviewed. He stated:

> If this attorney and his statistician were to ever be peer reviewed, please send that to me, all right? And "peer reviewed" generally means that other researchers have looked at this, and they have tried to replicate the study and that they've done it with the same level of success he's had.[103]

Peer review is part of the process used by scientists to determine which studies should be published in a given journal.[104] A researcher first conducts a study and submits a manuscript to the editor of a journal.[105] The editor then asks peers (who remain anonymous) to comment upon the manuscript and the methods used in the study.[106] The editor uses the feedback to decide if the manuscript should be accepted, rejected, or sent back to the author with instructions to revise and resubmit.[107] While peer approval does not guarantee that a study is reliable and valid, it is a useful proxy for quality (in the same way that price is a useful proxy for quality when buying clothing).[108]

Law reviews go through a selection and editing process, but most are not peer reviewed.[109] In scientific fields, peer-reviewed journals are often considered

103. *Hearing Transcript*, *supra* note 7, at 15.

104. Dale J. Benos et al., *The Ups and Downs of Peer Review*, 31 ADVANCES IN PHYSIOLOGY EDUC. 145, 145 (2007).

105. *Id*.

106. *Id*.

107. *Id*. at 145–46.

108. *See id.* at 148; ROBERT M. MILARDO, CRAFTING SCHOLARSHIP IN THE BEHAVIORAL AND SOCIAL SCIENCES: WRITING, REVIEWING, AND EDITING 148 (2015) (stating that "[s]cholars overwhelmingly believe peer review improves the quality of published papers . . . .").

109. *See generally* Carol Sanger, *Editing*, 82 GEO. L.J. 513 (1993). One of the two studies the judge criticized was, we believe, actually reviewed by a professor before the journal extended an offer of publication. E-mail from Shu-en Wee, Former Editor, COLUM. L. REV. ONLINE (July 11, 2017, 08:28 a.m. CST) ("your piece was reviewed by one professor before an offer was extended.") (on file with author); *see also Submission Instructions: Peer Review*, COLUM. L. REV., http://columbialawreview.org/submissions-instructions/ (last visited Nov. 21, 2017) (stating "[b]ecause peer review of articles and essays improves the *Columbia Law Review*'s selection process and helps to verify piece originality, the *Review* strongly prefers subjecting submitted pieces to peer review, contingent on piece–selection timeframes and other extenuating circumstances.").

to be more prestigious than non-peer-reviewed journals,[110] although the difference in quality may not be as large as the judge, and many others, think.

For example, one controlled study of the peer-review system found that reviewers detected only 25% of the errors that were intentionally inserted into a manuscript under consideration.[111] Another author lamented that "[t]he peer-reviewed articles with which I am most familiar all turned out to have severe methodological errors that were not identified . . . prior to publication."[112] Other weaknesses of the peer-review system include letterhead bias on the part of the reviewers and the journals' desire to publish extraordinary findings rather than "replication studies."[113]

While the value of peer review is debatable, the judge's specific concern reveals a misunderstanding of the process. When criticizing the two studies, he said: "'peer reviewed' generally means that other researchers . . . have tried to replicate the study and that they've done it with the same level of success he's had."[114] This is not accurate. Study replication is *not* part of the peer review process.[115] Peer reviewers merely read and evaluate the manuscript under consideration and offer comments; they do not replicate the study.[116]

Replication is one of the main ways that behavioral research progresses. Researchers have more faith in the reality of a phenomenon or the validity of a theory if the phenomenon has been observed on multiple occasions, or if a theory's predictions have been confirmed on multiple occasions in different settings.[117] This is why both of the studies criticized by the judge included a section that called for "further testing" by other researchers.[118]

But most significantly, the judge was unaware that the second of the two studies was, in fact, a replication of the first. As we explained in that study:

110. *See* Benos, *supra* note 104, at 146.

111. Ed Diener, *A Website System for Communicating Psychological Science*, 12 PERSP. ON PSYCHOL. SCI. 684, 684–85 (2017).

112. Steven Lubet, *Law Review vs. Peer Review: A Qualified Defense of Student Editors*, 2017 U. ILL. L. REV. 1, 3 (2017) (discussing several flawed peer-reviewed articles as well as drawbacks to the peer-review system).

113. Diener, *supra* note 111, at 685.

114. *Hearing Transcript*, *supra* note 7, at 15.

115. THOMAS H.P. GOULD, DO WE STILL NEED PEER REVIEW?: AN ARGUMENT FOR CHANGE 7 (2013).

116. *Id.*

117. *See* Stefan Schmidt, *Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences*, 13 REV. GEN. PSYCHOL. 90, 90–91 (2009) (discussing the importance of replication).

118. *See Empirical Test*, *supra* note 13, at 1159–60; *Conceptual Replication*, *supra* note 14, at 34–35; *see also Hearing Transcript*, *supra* note 7, at 13–15.

> [O]ur main objective is to test the reliability of our previous finding *by replicating the study*. In order to do this, we designed and conducted a conceptual replication rather than a direct replication. A conceptual replication retests the original hypothesis but intentionally varies specific features of the original methodology. A benefit of conceptual replication is that it addresses one of the common weaknesses of psychological research: limited generalizability . . .
>
> A conceptual replication allows us to address this limitation by testing our hypothesis under a different set of circumstances . . . [O]ur new study has a larger sample size, a different fact pattern, and includes stronger evidence of the defendant's guilt. We also provided mock jurors with a shorter underlying instruction on reasonable doubt. However, the variable being tested—the mandate "not to search for doubt" but instead "to search for the truth"—is the identical language that we tested in our previous study.[119]

Findings in behavioral research that have been successfully replicated deserve special status because not all replication attempts are successful. In fact, in a recent, large-scale, multi-site examination of peer-reviewed studies, more than 60% of the studies tested were not replicated.[120] This is why scientists, policymakers, and others increasingly rely on meta-analysis, a statistical procedure that combines the results of multiple studies to draw evidence-based conclusions.[121]

III. "THE MOTHER OF ALL BIASES"

The previous Part identified and debunked seven different criticisms of the two behavioral studies. A possible explanation for the judge's misplaced concerns is confirmation bias. Confirmation bias is the automatic human tendency to seek and recall evidence that confirms one's belief and, at the same time, fail to notice or recall evidence that disconfirms one's belief.[122]

It appears that the judge first decided to deny the defense lawyer's motion and then tried to leverage support to confirm his reasoning. For example, when the defendant's lawyer said, correctly, "it is a controlled study," the judge countered, "[n]o, it's not."[123] Moreover, the judge claimed to have thoroughly

119. *Conceptual Replication*, *supra* note 14, at 27 (emphasis added).

120. Benedict Carey, *Many Psychology Findings Not as Strong as Claimed, Study Says*, N.Y. TIMES, Aug. 28, 2015, at A1.

121. ARON & ARON, *supra* note 68, at 258–59.

122. *See* THOMAS GILOVICH & LEE ROSS, THE WISEST ONE IN THE ROOM: HOW YOU CAN BENEFIT FROM SOCIAL PSYCHOLOGY'S MOST POWERFUL INSIGHTS 138–40 (2015).

123. *Hearing Transcript*, *supra* note 7, at 14.

analyzed the studies—"[b]elieve me, I researched the heck out of them"[124]—yet, apparently he failed to notice multiple aspects of the study: first, that the study participants were not college students; second, that all *p*-values were significant or highly significant; and third, that the second study was a replication of the first.

Readers may be inclined to think that this judge is an outlier, but the evidence suggests that this problem is more widespread. Scholars have observed that, in many cases, "judges have made use of social scientific evidence only when it was supportive of the ruling a judge wanted to make anyway. And, sometimes, the courts have ignored, dismissed, or misrepresented the findings of social scientific research."[125]

Such confirmation bias is pervasive. Quite accurately, psychologist Scott Lilienfeld has called it "the mother of all biases."[126] The solution to confirmation bias "is easy to state but difficult to follow."[127] Judges must slow down and deliberately look for information that challenges their preconceptions and prior beliefs. They must be their own devil's advocate.

On an even more fundamental level, confirmation bias raises this troubling question: Do judges *want* to make good use of behavioral research? We are fearful that, for some judges, the answer is no. For example, near the end of the motion hearing, the judge made an astonishing statement: "Frankly, Mr. [defense lawyer], I think you can just ask [to modify the jury instruction] without going through the statistical stuff, I would probably be more inclined to grant it."[128]

The judge essentially stated that he might have granted the motion if the defense lawyer had not supplemented it with two published studies. It is clearly illogical to assert that an argument has merit *per se* but will be rejected because the meritorious argument *is also supported by empirical data*. Even if the studies had contained some methodological weaknesses—if the *p*-values had been too high, for example, or if the study participants had all been 20-year-old college students—none of that should cause a judge to pivot 180-degrees and deny a motion he would otherwise be inclined to grant.

It is difficult to explain the judge's thinking. Perhaps he wished to send a strong signal to lawyers who appear before him: "Your motions will be more likely to prevail if you stick to legal argument and do not present empirical evidence." If this is the case, his motivation could be explained as follows:

124. *Id*. at 19.
125. COSTANZO & KRAUSS, *supra* note 33, at 24.
126. GILOVICH & ROSS, *supra* note 122, at 140.
127. *Id*. at 147.
128. *Hearing Transcript*, *supra* note 7, at 18.

> Intellectually, judges know little about empirical research and are unable (or perhaps unwilling) to make sense of it. . . . But the resistance is not only intellectual. There are also personal reasons behind the reluctance of judges. Judges tend to be self–confident, politically conservative, and protective of their prestige and power. When confronted with empirical research, they are likely to feel that they do not need help from social scientists; they are likely to suspect that social scientists are politically liberal, and they may view social science as undermining their power.[129]

The problem with such judicial thinking, however, is behavioral scientists sometimes discover inconvenient truths that disrupt the status quo and require procedural reforms. Once again, the words of Oliver Wendell Holmes, Jr., continue to ring true well over a century after he penned them:

> It is revolting to have no better reason for a rule of law than that . . . it was laid down in the time of Henry IV. It is still more revolting if the grounds upon which it was laid down have vanished long since, and the rule simply persists from blind imitation of the past.[130]

Unfortunately, as discussed above, many judges—as well as many lawyers—have little or no appetite for change. But for those who are not averse to change and are willing to consider evidence in support of it, the question remains: How should behavioral research be evaluated in court?

IV. GRADING THE VALIDITY OF THE STUDIES

The overall quality of behavioral research is determined by four kinds of validity: internal validity,[131] construct validity,[132] statistical conclusion validity,[133] and external validity.[134]

"Internal validity" refers to the degree to which we can be confident there is a causal relationship between two variables.[135] For example, if variables X and Y are correlated, that does not necessarily mean that X caused Y. It is quite possible that a third variable caused both X and Y.

To assess internal validity, behavioral scientists ask questions like these: (1) Did the researcher randomly assign participants to various versions of the *test*

129. COSTANZO & KRAUSS, *supra* note 33, at 24–25.

130. Oliver Wendell Holmes, Jr., *supra* note 1, at 469.

131. THOMAS D. COOK & DONALD T. CAMPBELL, QUASI-EXPERIMENTATION: DESIGN & ANALYSIS ISSUES FOR FIELD SETTINGS 37 (1979).

132. *Id.* at 38.

133. *Id.* at 37.

134. *Id.*

135. COOK & CAMPBELL, *supra* note 131, at 38.

variable—for example, different reasonable-doubt instructions—and then measure each participant's performance on the *outcome variable*—for example, juror verdict? (2) Did the researcher hold all other variables—for example, case facts—constant to avoid influencing the scores on the outcome variable? (3) Did the researcher take steps to eliminate experimenter bias by limiting his or her interaction with the study participants?

"Construct validity" refers to the adequacy of the operational definitions used by the researcher.[136] An operational definition is the specific way in which a researcher realizes or measures a variable.[137] For example, intelligence can be measured directly with a standardized IQ test of cognitive abilities in several domains, or it can be measured indirectly via school grades. The first measure has more construct validity than the second measure.

To assess construct validity in the context of jury-instruction studies, behavioral scientists ask questions like these: (1) Was the test variable—that is, different versions of reasonable-doubt instructions—defined and presented in a precise manner that can be replicated by other researchers? (2) Was the outcome variable—that is, juror verdicts—measured in way that is both reliable (reproducible) and valid (accurate)?

"Statistical conclusion validity" refers to the reliability and accuracy of a study's statistical conclusions; it also refers to a study's ability to identify statistical relationships and effects that are weak but real.[138] Error rates are higher in some studies than others, and some studies are underpowered, as they include too few observations to detect small effects.

To assess statistical conclusion validity, behavioral scientists ask questions like these: (1) Did the study include a sufficiently large number of participants?[139] (2) Did the researcher use appropriate statistical procedures?[140] (3) Did the researcher set the probability of a Type I error[141] at a suitably low level, usually 5%?

136. *Id*. at 38.

137. *Variables and Operational Definitions*, CERT, https://cirt.gcu.edu/research/developmentresources/research_ready/quantresearch/variables_def (last visited Jan. 23, 2018).

138. *See generally* COOK & CAMPBELL, *supra* note 131, at 39–41.

139. Large samples produce more accurate findings than smaller samples, *ceteris paribus*. They also are more likely to detect small effects.

140. For example, when the outcome variable is dichotomous (guilty or not guilty) instead of scaled (degree of guilt on a 10-point scale), a different procedure must be used to analyze the impact of the different jury instructions.

141. A Type I error is a false positive; it occurs when a researcher concludes that the test variable had an impact on the outcome variable but, in reality, it did not. ARON & ARON, *supra* note 68, at 261.

"External validity" refers to the degree to which a study's findings can be generalized (applied) to other persons and settings.[142] For example, if a researcher observes that reducing the number of inmates in a particular jail leads to fewer disciplinary problems per inmate, will the same effect be observed in other jails? If it is, the finding is said to have a high degree of external validity.

To assess external validity, behavioral scientists ask questions like these: (1) To what degree are the study's participants representative of the larger population to which the researcher wishes to generalize? (2) To what extent did features of the study approximate or mimic features in the real world? (3) Have the study's findings been replicated with a different group of participants?

These four validities do not operate independently of one another. That is, when researchers take steps to strengthen one kind of validity, another kind of validity may be weakened. For example, highly controlled conditions increase a study's internal validity, but such conditions are often artificial, and artificiality decreases a study's external validity. For this reason, it is rare to find a single study that exhibits high degrees of validity on all four dimensions. In most areas of scientific research, internal validity is the most desirable of the four.[143]

How did the two behavioral studies discussed throughout this Article fare in terms of the four validities? First, both studies were experiments that used random assignment and procedural controls to eliminate the influence of extraneous variables and experimenter bias. These features are associated with high internal validity.

Second, in both studies, the test variable (a particular version of the reasonable-doubt instruction) and the outcome measure (a mock juror's verdict) were operationally defined in a precise and legally appropriate manner. These features are associated with high construct validity.

Third, both studies used large numbers of participants—more, in fact, than other reasonable-doubt studies—and set the Type I error rate at 5%. These features are associated with high statistical conclusion validity.

Fourth, neither study occurred in an actual judicial setting; rather, as discussed earlier, both studies used the case summary method. However, both studies recruited participants from diverse backgrounds and used realistic case materials, and the second study successfully replicated the results of the first study. These features are associated with at least a moderate degree of external validity.

In summary, the defense lawyer who filed the motion was correct: the two studies provided reliable and valid evidence in support of his request to modify the pattern jury instruction on the burden of proof.

142. COOK & CAMPBELL, *supra* note 131, at 70–71.

143. *Id.* at 83.

CONCLUSION

Behavioral research has played, and will continue to play, a significant role in the legal system, including in criminal jury trials. Therefore, judges and lawyers must educate themselves in the fundamentals of scientific and statistical reasoning. In this Article we have used a criminal-case motion hearing as a "case study" to identify and debunk common misconceptions about behavioral research.

To begin, judges and lawyers should avoid *ad hominem* attacks. Behavioral research must be evaluated on its merits and cannot be dismissed because of the perceived personal faults of the researchers.[144] Further, when evaluating the research itself, it is important to be able to distinguish between surveys, which seek to estimate the frequency of a specific characteristic in the larger population, and controlled studies, which are designed to detect differences *between* test groups resulting from a manipulated variable, such as a jury instruction.[145]

When criticizing a controlled study, judges should articulate a basis for their criticisms. It should not be sufficient to simply dismiss statistical studies when they do not align with personal beliefs on an issue.[146] In particular, regarding statistics, judges and lawyers must also take time to comprehend that, while the sample size of a controlled study is important, it is not the same as statistical significance. Rather, the sizes of the groups being tested, when considered along with the observed effect size, can be used to determine the statistical significance of a study's findings as expressed by the *p*-value.[147]

When evaluating the design of a controlled study, judges and lawyers must be able to distinguish between the various forms of bias for which researchers should control, and the so-called "bias" of test participants which is eliminated by randomly assigning participants to the test groups.[148] Also with regard to study design, the lack of juror deliberations does not invalidate a study. In fact, most published studies of jury instructions do not employ deliberations, and those that do have failed to demonstrate any systematic effect on verdicts.[149]

Furthermore, with regard to study evaluation, judges and lawyers must understand the concept of peer review. It is true that *most* law reviews are not peer reviewed. However, peer review is just that: review and comments by a peer or peers solicited by a journal before it extends an offer to publish an article. Peer review does not guarantee a study's quality; in fact, controlled experiments of

144. *See supra* Part II.A.
145. *See supra* Part II.B.
146. *See supra* Part II.C.
147. *See supra* Part II.D.
148. *See supra* Part II.E.
149. *See supra* Part II.F.

the peer-review process itself have exposed several flaws and failures. And most significantly, peer reviewers do *not* replicate a study. Study replication is an entirely different process and, for several reasons, is very desirable but also rare.[150]

Confirmation bias is often the cause of the kinds of erroneous thinking discussed and debunked in this Article. That is, judges sometimes first make up their mind and then look for things that support their beliefs. Additionally, judges may misinterpret, or even misrepresent, statistical findings in order to sustain their predetermined conclusions.[151]

Finally, in order to guard against confirmation bias and to properly evaluate behavioral research findings, we have described the four kinds of validity that behavioral scientists use to evaluate controlled studies. We have also suggested specific questions that judges and lawyers can ask during the evaluation process.[152]

Understanding the basic concepts discussed in this article is an important first step if judges and lawyers wish to make good use of behavioral-science evidence.

150. *See supra* Part II.G.

151. *See supra* Part III.

152. *See supra* Part IV.