MEASURING FAIRNESS

Barry C. Edwards, J.D., Ph.D.

MEASURING FAIRNESS

*Barry C. Edwards, J.D., Ph.D.**

Courts are tasked with determining whether criminal defendants received fair trials yet currently lack a rigorous method for assessing fairness. To address this problem, this Article outlines a valid and reliable measure of trial fairness grounded in feasible survey research. A key element of this method is identifying the relationship between the verdict preferences in a jury pool and the probability that juries selected from the pool will render a guilty verdict. Identifying this relationship enables estimation of outcome probabilities that would otherwise be unattainable. The author demonstrates this method by applying it to real criminal trials, successfully classifying test cases involving ineffective assistance of counsel, coerced confessions, improper prosecutorial arguments, and flawed jury instructions. To ensure transparency and accessibility for legal and social science audiences, the author details how fairness can be measured and offers the analytic tools in an open-source statistical framework.

INTRODUCTION

At present, one of the most fundamental legal questions—did the defendant receive a fair trial?—cannot be answered objectively. As an abstract legal concept, fairness has been clearly defined: A trial is considered fair if its errors and deficiencies do not create a "reasonable probability of a different outcome."[1] Unfortunately, courts lack a clear method of applying this standard to determine whether defendants received fair trials. Current approaches do not permit courts to quantitatively assess the extent to which inadmissible evidence, improper arguments, flawed instructions, or any other trial errors increased the probability of a guilty verdict.[2]

The inability to quantify fairness is deeply troubling given the fundamental legal principle that trials should be fair.[3] If fairness cannot be measured, we should not expect trials to be fair.[4] General statements about the importance of

* Barry C. Edwards is the founder and director of Fair Trial Analysis, a non-profit organization that supports the right to a fair trial with research. He received his B.A. from Stanford University, a J.D. from New York University, and a Ph.D. from the University of Georgia. He is co-author of a popular series of political science research methods textbooks and continues to teach part-time for the University of Georgia.

1. *See* Strickland v. Washington, 466 U.S. 668, 694 (1984). Part I discusses the development and scope of the *Strickland* fairness standard.

2. Past cases provide some guidance on the effect of an error, but every case is different. Academic studies can quantify how some item of evidence, argument, or instruction affects typical jurors, but their stylized fact patterns do not fit real trials. *See infra* Part II.B.

3. Fairness is a bedrock principle of the legal system. It is not a partisan issue or something only defendants want. The U.S. Supreme Court has appropriately recognized that "[s]ociety wins not only when the guilty are convicted but when criminal trials are fair; our system of the administration of justice suffers when any accused is treated unfairly." Brady v. Maryland, 373 U.S. 83, 87 (1963). It should not be necessary to belabor the assumption that trials should be fair.

4. Peter Drucker's proverbial message about measurement is fitting: "What gets measured gets improved." *E.g.*, *What Gets Measured Gets Improved*, VERINT (July 23, 2013), https://www.verint.com/blog/what-gets-measured-gets-improved/ [https://perma.cc/4A8D-6TUZ]; *see also* Paul Zak, *Measurement Myopia*, THE DRUCKER INST. (July 4, 2013), [https://perma.cc/Y87G-HC7L] ("If you can't measure it, you can't manage it."). Although Peter Drucker has written extensively on the importance of measurement to management, these specific quotations, widely attributed to Drucker, may be

fairness, justice, and the rule of law will not suffice. To fulfill the promise of fair trials for all defendants, there must be a valid, reliable, and efficient method to measure fairness. This Article delivers good news: The fairness of trials can be measured effectively using standard methodologies and readily obtained data, enabling accurate and efficient differentiation between fair and unfair trials.[5] The Article provides a detailed explanation of how fairness can be measured and demonstrates the method's validity and reliability with empirical data and a series of test cases.

Part I presents an overview of the "reasonable probability of a different outcome" standard, a benchmark used to judge trial fairness in many contexts.[6] Although this standard offers a clear conceptual definition of fairness, no effective method has been developed to objectively assess potential trial outcomes. Even with a precise understanding of a trial error and its impact on jurors, estimating the probability of a different trial outcome remains impossible.[7] This inability to quantify the probability of a different trial outcome undermines the right to a fair trial and benefits no one.[8]

Part II outlines a method of estimating the probability of a different outcome using readily obtainable data. This process begins with case-specific estimates of verdict preferences in a jury pool, obtained through careful survey research—a common form of analysis that, without context, is insufficient for predicting final jury verdicts. To bridge this gap, Part II introduces an elegant

---

paraphrases rather than direct quotations. Zak, *supra*. Famed physicist William Thomson (better known as Lord Kelvin) offered similar advice about measuring quantities of interest:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.

1 WILLIAM THOMSON, *Electrical Units of Measurement*, *in* POPULAR LECTURES AND ADDRESSES 73, 73–74 (London, MacMillan & Co. 1889).

5. We can solve the riddle of harmless error in some criminal appeals. This is not to say the riddle can be solved in all cases. Some results will be indeterminate due to estimation uncertainty. Additionally, Part IV identifies some types of criminal trials where it would still be difficult to objectively estimate the effect of a trial error. *See infra* Part IV.

6. This Article focuses on measuring the fairness of criminal trials, but some variation of the method may be used to measure the fairness of civil trials. *See infra* Part IV. Civil trial errors are also evaluated with harmless error analysis. *See* Robert M. Gross & David R. Maass, *Harmless Error in Civil Appeals*, 89 FLA. B.J. 10 (2015); *see also* Little Sisters of the Poor Saints Peter & Paul Home v. Pennsylvania, 591 U.S. 657, 684 (2020) (explaining that the rule of prejudicial error is a harmless error rule for civil appeals).

7. Even if it can be shown that a specific trial error increased the proportion of jurors who prefer a guilty verdict from, say, sixty percent to seventy percent, the error's effect on the probability of conviction still cannot be determined. A modest change in individual-level verdict preferences can have a substantial effect in a close case. A substantial shift in preferences may have only a modest effect on the probability of the fairness of a defendant's trial, and cannot be addressed objectively, even with perfect information about juror verdict preferences.

8. The prosecutor cannot prove that an error was harmless. Some judges may think an error was harmless, but other judges may have different opinions, leaving the analysis open to interpretation. The defendant also cannot prove that an error was harmful; even if a reviewing court thinks an error was harmful, that assessment is an opinion, rather than a scientific fact.

jury deliberation model, synthesizing decades of empirical research to define the relationship between jurors' initial verdict preferences and the probability of a guilty verdict.[9] Identifying the relationship between juror preferences and jury verdicts enables us to contextualize survey data and accurately estimate the probability of a different trial outcome.

Part III applies the fairness measurement method to real criminal trials, including cases studied by the author and others. The test cases in Part III address issues such as mitigation evidence omitted by ineffective defense counsel, coerced confessions, improper prosecutorial arguments, and the significance of jury instructions. Some of these trials were ultimately deemed fair, while others were judged unfair. Applied to real criminal trials, the analytical method developed in this Article consistently replicates appellate court judgments. Part IV explores potential avenues for future research.

The potential benefit of a valid and reliable test of fairness may be comparable to the contribution of forensic tests in criminal trials.[10] Just as biological samples are routinely analyzed and interpreted to address questions of guilt and innocence, survey samples could be analyzed and interpreted to address questions of fairness. A valid and reliable fairness measure would enable appellate courts to evaluate criminal trials in a clear, evidence-based framework and determine whether retrials are necessary. Like forensic DNA testing, which has been effective in convicting the guilty and exonerating the innocent,[11] a fairness measure could help courts uphold fair trials and undo unfair trials. An objective fairness measure would reduce reliance on subjective judgments by providing a structured, quantifiable method for analyzing fairness. To ensure transparency and accessibility, the core deliberation model and associated

9. The method uses a fair amount of math and statistics to identify the relationship between juror preferences and jury verdicts as precisely as possible. Rest assured, the method does not require manually performing the calculations discussed in this Article; they are completed by computers. The author discusses underlying math equations and statistical formulas to make the technical aspects of this analysis transparent and visible.

10. *See generally* Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCI. 892 (2005) (describing DNA tests supplanting older identification tools like blood typing, hair comparison, and fingerprint matching). The effect of a fairness measure will be smaller than that of DNA testing for several reasons. Only a fraction of criminal investigations result in trials, and only a small subset of trial verdicts are challenged on appeal. *See generally* Gregory M. Gilchrist, *Trial Bargaining*, 101 IOWA L. REV. 609 (2016) (explaining the effects of plea bargaining on criminal trials and appeals). At the same time, appellate decisions about the procedural fairness affect plea bargaining and decisions to appeal. *See generally* Michael M. O'Hear, *Plea Bargaining and Procedural Justice*, 42 GA. L. REV. 407 (2008) (applying the procedural justice framework to the plea bargaining and appellate process).

11. *See generally* JAY ARONSON, GENETIC WITNESS: SCIENCE, LAW, AND CONTROVERSY IN THE MAKING OF DNA PROFILING (2007) (detailing the road to acceptance of DNA evidence in courtroom trials); Brandon L. Garrett, *Judging Innocence*, 108 COLUM. L. REV. 55 (2008) (analyzing case histories of first 200 people exonerated by post-conviction DNA testing); EDWARD CONNORS ET AL., CONVICTED BY JURIES, EXONERATED BY SCIENCE (1996) (cataloguing 28 cases where DNA evidence exonerated criminal defendants); JIM DWYER, PETER NEUFELD & BARRY SCHECK, ACTUAL INNOCENCE (2000) (narrating stories of several convicted defendants later exonerated by DNA evidence).

functions for analyzing the harmfulness of trial errors are freely available in an open-source statistical software package.[12]

I. REASONABLE PROBABILITY OF A DIFFERENT RESULT AS A STANDARD OF FAIRNESS

Deciding whether a defendant's trial was fair requires comparing the trial the defendant received to a hypothetical, error-free version of the defendant's trial.[13] In many situations, courts compare the defendant's actual trial to a hypothetical "perfect" trial and ask whether there is "reasonable probability of a different outcome" between them.[14] If there is a reasonable probability of a different outcome, the defendant did not receive a fair trial and is entitled to relief.[15] If there is little or no probability of a different outcome, the trial errors were harmless, the defendant's trial was fair enough, and his conviction should stand.[16]

Courts apply harmless error analysis and the "reasonable probability of a different outcome" standard to a wide variety of claims.[17] The harmless error

12. Technical discussion of the software package and its analytic routines are found in footnotes. The SATE Package is a library for the freely available, open-source R program for statistical computing. The package includes detailed explanations of analytic routines with examples of code usage. *See* Barry Edwards, *SATE: Scientific Analysis of Trial Errors*, THE COMPREHENSIVE R ARCHIVE NETWORK (2025), https://cran.r-project.org/package=sate [https://perma.cc/W4C2-EL9J] [hereinafter *SATE*].

13. *See* Barry C. Edwards, *A Scientific Framework for Analyzing the Harmfulness of Trial Errors*, 8 UCLA CRIM. JUST. L. REV. 1, 18 n.89 (2024) [hereinafter *A Scientific Framework*].

14. The Supreme Court has occasionally suggested focusing on the harmfulness of errors in the defendant's actual trial without making comparisons to hypothetical trial conditions. *See, e.g.*, Sullivan v. Louisiana, 508 U.S. 275, 279–80 (1993). It is, however, impossible to evaluate the effect of an error or omission without making comparisons. *See generally* Donald B. Rubin, *Causal Inference Using Potential Outcomes: Design, Modeling, Decisions*, 100 J. AM. STAT. ASSOC. 322 (2005) (surveying several approaches and models used to infer causal effects); Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17 (2011) (reviewing advances made towards credible causal inference in empirical legal studies). The impossibility of focusing on a jury's "actual reliance" on wrongly admitted evidence is most obvious in post-conviction proceedings about the effect of evidence the jury never heard due to ineffective assistance of counsel or prosecutorial misconduct. *See* Edwards, *A Scientific Framework*, *supra* note 13, at 29.

15. A trial error's potential effect on a criminal defendant is evident, but other harms may also be considered, including its impact on the jury's understanding of the evidence. There are multiple approaches to assessing the harmfulness of trial errors. *See, e.g.*, Jason M. Solomon, *Causing Constitutional Harm: How Tort Law Can Help Determine Harmless Error in Criminal Trials*, 99 NW. U.L. REV. 1053 (2005) (arguing for judges to follow a hybrid approach to determine causation in criminal cases); Jeffrey O. Cooper, *Searching for Harmlessness: Method and Madness in the Supreme Court's Harmless Constitutional Error Doctrine*, 50 KAN. L. REV. 309 (2002) (arguing that the Supreme Court's harmless error doctrine has run amok because the Court is not recognizing legal fictions that underlie the doctrine); Stuart P. Green, *The Challenge of Harmless Error*, 59 LA. L. REV. 1101 (1999) (evaluating the harmless error doctrine through the ethical frameworks of consequentialism and non-consequentialism).

16. The general rule can be stated simply, but the specific standards governing criminal appeals, such as the burden of proof and certainty required, depend on the procedural context and the claims being litigated. For example, a prisoner challenging a state court conviction in federal court will be required to demonstrate greater error than he will in post-conviction proceedings in state court. *See, e.g.*, Shinn v. Ramirez, 596 U.S. 366, 377 (2022).

17. According to Professor Landes and Judge Posner, the harmless error doctrine is "probably the most cited rule in modern criminal appeals." William M. Landes & Richard A. Posner, *Harmless Error*, 30 J.

rule was first applied to ordinary trial errors, which are mistakes judges make during trials that do not implicate a defendant's constitutional rights.[18] The U.S. Supreme Court's *Chapman v. California* decision extended harmless error analysis to constitutional trial errors, such as a prosecutor commenting on a defendant's decision not to testify.[19] In *Chapman*, the Court asked whether there was a "reasonable possibility" that improper comments contributed to the defendant's conviction.[20]

Although defendant Chapman won her appeal, the *Chapman* decision was not a victory for criminal defendants. *Chapman* set a precedent for applying harmless error analysis to constitutional trial errors, thereby expanding the scope of trial errors appellate courts can excuse as harmless.[21] Following the logic of *Chapman*, the Court has applied harmless error analysis to other constitutional trial errors, including trials with coerced confessions,[22] illegally obtained evidence,[23] and violations of defendants' right to confront witnesses.[24] According to the Court, in direct appeals, a trial error should "be quantitatively assessed in the context of other evidence presented in order to determine whether its admission was harmless beyond a reasonable doubt."[25] The prosecution must demonstrate that errors in the defendant's trial were

LEGAL STUD. 161, 161 (2001). Similarly, Judge Walker writes that the harmless error rule determines the fate of "more criminal appeals than any other doctrine." John M. Walker Jr., *Harmless Error Review in the Second Circuit*, 63 BROOK. L. REV. 395, 395 (1997). As Lee Kovarsky recently observed, "These rules are everywhere. Few criminal procedure inquiries are as omnipresent as the harmless-error doctrine." Lee Kovarsky, *Outcome Sensitivity and the Constitutional Law of Criminal Procedure*, 98 IND. L.J. 429, 430 (2023).

18. Ordinary trial errors typically concern rules of evidence, criminal procedure, and jury instructions. For examples of ordinary, non-constitutional trial errors, see Molina-Martinez v. United States, 578 U.S. 189, 194, 197–98 (2016) (requiring a showing of reasonable probability of different outcome but for incorrect application of sentencing guidelines); Greer v. United States, 593 U.S. 503, 507–08 (2021) (requiring a showing of reasonable probability of different outcome but for mistaken jury instructions); Neder v. United States, 527 U.S. 1, 8–10 (1999) (applying harmless error analysis to the omission of an element of the crime in jury instructions).

19. Chapman v. California, 386 U.S. 18, 24–25 (1967). It is important to note that structural errors are treated differently from trial errors. Appeals courts continue to grapple with the distinction between structural errors, which warrant automatic reversal, and trial errors, which require harmless error analysis. *See* Weaver v. Massachusetts, 582 U.S. 286, 294–97 (2017). For recent commentary on the distinction, see generally Zachary L. Henderson, *A Comprehensive Consideration of the Structural-Error Doctrine*, 85 MO. L. REV. 965 (2020) (describing origins of the structural-error doctrine and surveying structural errors identified by the Supreme Court and circuit courts); Susan Yorke, *Jury Nullification Instructions as Structural Error*, 95 WASH. L. REV. 1441 (2020) (arguing that anti-nullification jury instructions are ill-suited to harmless error analysis); *Scientific Framework*, *supra* note 13, at 4–6 (enumerating structural errors identified by the Supreme Court).

20. *Chapman*, 386 U.S. at 23.

21. *See* Gregory Mitchell, *Against "Overwhelming" Appellate Activism: Constraining Harmless Error Review*, 82 CAL. L. REV. 1335, 1335–36 (1994).

22. Arizona v. Fulminante, 499 U.S. 279, 306–12 (1991); *see also* Charles J. Ogletree Jr., Arizona v. Fulminante*: The Harm of Applying Harmless Error to Coerced Confessions*, 105 HARV. L. REV. 152, 155–56 (1991) (discussing *Fulminante* and the common law roots of the harmless error rule).

23. Chambers v. Maroney, 399 U.S. 42, 52–53 (1970).

24. Harrington v. California, 395 U.S. 250, 252–54 (1969); Delaware v. Van Arsdall, 475 U.S. 673, 674 (1986).

25. *Fulminante*, 499 U.S. at 308.

harmless.[26] That is, the prosecution bears the burden of proving that errors in the defendant's trial did not cause a reasonable probability of a different outcome.[27]

The "reasonable probability of a different outcome" standard is also used to evaluate errors of omission in post-conviction proceedings. In *Strickland v. Washington*, the Court applied the standard to assess the harmfulness of ineffective counsel failing to present mitigation evidence during capital sentencing.[28] Similarly, the standard is used to evaluate a prosecutor's failure to disclose exculpatory evidence.[29] The materiality of evidence withheld in violation of the Constitution hinges on its likely effect.[30] The court must determine whether there is a reasonable probability that, if the evidence had been disclosed, the result of the trial would have been different.[31]

There is yet one more significant category of cases where the probability of a different outcome is a central issue.[32] When defendants seek new trials based on newly discovered evidence, courts assess whether the new evidence would likely produce a different result if a new trial were granted.[33] Federal courts have not been receptive to post-conviction claims based on newly discovered evidence.[34] However, some state laws entitle defendants to relief if there is a reasonable probability that the new evidence would have changed the outcome had it been available during the original trial.[35] Some states, including New

26. Chapman v. California, 386 U.S. 18, 23 (1967). The difference in the probability of conviction may also be relevant to certain pretrial motions, such as a motion to change the trial venue due to pretrial publicity and a motion in limine to exclude unduly prejudicial evidence.

27. *Id.*

28. Strickland v. Washington, 466 U.S. 668, 694–96 (1984).

29. Brady v. Maryland, 373 U.S. 83, 88–91 (1963).

30. *See* Kyles v. Whitley, 514 U.S. 419, 440–41 (1995) (considering the cumulative effect of evidence withheld by the prosecutor under the reasonable probability of a different outcome standard); Strickler v. Greene, 527 U.S. 263, 289–91 (1999) (considering whether the suppression of witness impeachment evidence would cause a reasonable probability of a different outcome); Cone v. Bell, 556 U.S. 449, 452 (2009) (considering whether a suppressed statement supporting insanity defense would create a reasonable probability of a different sentence).

31. *See* United States v. Bagley, 473 U.S. 667, 682 (1985) ("The evidence is material only if there is a reasonable probability that, had the evidence been disclosed to the defense, the result of the proceeding would have been different.").

32. There are specific situations where courts use the reasonable probability of a different outcome standard but focus on outcomes other than jury verdicts. For example, in cases where post-conviction petitioners claim that they mistakenly accepted or rejected plea deals due to ineffective assistance of counsel, the petitioner must show that there is a reasonable probability that he would not have accepted or rejected the plea deal but for the ineffective assistance of counsel. *See, e.g.*, Missouri v. Frye, 566 U.S. 134, 147 (2012); United States v. Dominguez Benitez, 542 U.S. 74, 76 (2004).

33. *Kyles*, 514 U.S. at 434. DNA testing of physical evidence is the most well-known example, but newly discovered evidence may be a recanting witness, a previously unknown witness, the discovery of official corruption, or the debunking of junk science introduced at the defendant's trial.

34. *See* Herrera v. Collins, 506 U.S. 390, 399–401 (1993) (holding that actual innocence is not a constitutional claim for habeas relief).

35. In Wisconsin, for example, when a petitioner makes a prima facie case for a new trial based on newly discovered evidence, the court must then determine "whether a reasonable probability exists that a different result would be reached in a trial." State v. Armstrong, 700 N.W.2d 98, 130 (Wis. 2005); *see also* State

York,[36] Mississippi,[37] Utah,[38] and Vermont,[39] will grant post-conviction relief if petitioners can demonstrate that there is a "reasonable probability" of a different result from newly discovered DNA evidence. Other states require petitioners to show that their newly discovered evidence creates more than a "reasonable probability" of a different outcome; the petitioner may be required to show that the newly discovered evidence probably changes the result, or more likely than not changes the result.[40]

The "reasonable probability of a different outcome" standard governs a wide range of issues. Although it provides an adequate conceptual definition of fairness, there has been no way to quantify the probability of a different outcome, making it impossible to apply the standard to real litigation.[41]

II. ESTIMATING THE PROBABILITY OF A DIFFERENT TRIAL RESULT

This Part explains how to measure the fairness of a criminal trial using a few data points that can be obtained from simple surveys. It discusses a method of measuring fairness in terms of the probability of a different result in a hypothetical, error-free trial. The reader is taken under the hood, so to speak, and offered a detailed account of the method.[42]

This Part introduces specific terminology. P(g) represents the probability that a juror randomly selected from the jury pool would vote for a guilty verdict. P(G) represents the probability that a jury returns a guilty verdict. The use of lowercase "g" and uppercase "G" signifies the difference between individual-level and aggregate-level outcomes. P(g | actual) is the probability that a random juror from the pool would vote for a guilty verdict given the actual trial condition. P(g | hypo) represents that probability in the hypothetical trial condition. The | symbol represents conditional probabilities. P(G | actual) and

v. Plude, 750 N.W.2d 42, 56 (Wis. 2008) (granting a new trial after the petitioner discovered that the prosecution's medical expert lied about his credentials at trial).

36. N.Y. CRIM. PROC. LAW § 440.10(1)(g-1) (McKinney 2025).

37. MISS. CODE ANN. § 99-39-5(1)(f) (West 2025).

38. UTAH CODE ANN. § 78B-9-104(1)(f) (West 2025).

39. VT. STAT. ANN. tit. 13 § 5566(a)(1) (West 2025).

40. For a thorough review of state laws, see Justin Brooks, Alexander Simpson & Paige Kenab, *If Hindsight is 20/20, Our Justice System Should Not Be Blind to New Evidence of Innocence: A Survey of Post-Conviction New Evidence Statutes and a Proposed Model*, 79 ALB. L. REV. 1045, 1058–66 (2015).

41. *See Scientific Framework*, *supra* note 13, at 26–33 (discussing courts' inability to estimate the harmfulness of trial errors); Solomon, *supra* note 15, at 1064, 67–68 (reporting that less than 20% of published federal appellate opinions on the harmfulness of error used a test for assessing harm and concluding the process is "largely unguided"); Dennis J Sweeney, *An Analysis of Harmless Error in Washington: A Principled Process*, 31 GONZ. L. REV. 277, 280 (1995) (asserting that harmless error analysis is currently "an exercise in judicial speculation").

42. For the results of analyzing the fairness of real criminal trials, see *infra* Part III.

P(G | hypo) are the conditional probabilities of primary interest because they allow us to quantify the probability of a different trial outcome.[43]

The process of estimating the probability of a guilty verdict can be broken down into four steps. First, one estimates verdict preferences in the jury pool using public opinion survey research.[44] Next, one calculates probabilities of each possible initial poll on juries selected from the jury pool based on a statistical rule.[45] Then, the probability of a guilty verdict starting from each possible initial poll is determined with a model of the deliberation process.[46] Finally, based on the vote and verdict probabilities from prior steps, one computes the overall probability of a guilty verdict, including standard errors of the estimates, and presents the results.[47]

To estimate the probability of a different trial outcome, one compares the probability of a guilty verdict in the actual and hypothetical trial conditions. Given values P(g | actual) and P(g | hypo), the four-step process yields corresponding values of P(G | actual) and P(G | hypo). The difference between P(G | actual) and P(G | hypo) is a measure of unfairness, and it can be used to determine whether a defendant received a fair trial.[48]

A. Estimating Verdict Preferences of a Jury Pool

The quantities P(g | actual) and P(g | hypo) should be unbiased estimates of a jury pool's verdict preferences in defendant's actual and hypothetical, error-free trial conditions.[49] Accurate estimates of P(g) are crucial because these estimates effectively summarize all the myriad factors that influence juror

43. P(G | actual) and P(G | hypo) could be estimated from a large sample of mock juries, but conducting hundreds or thousands of mock juries is not a feasible measurement strategy. Litigators rarely conduct even one mock jury due to the time and expense involved. *See generally* April J. Ferguson, *Selecting the Best Form of Jury Research for your Case & Budget*, OPVEON, https://www.opveon.com/blog/selecting-the-best-form-of-jury-research-for-your-case-budget? [https://perma.cc/Y8H4-6EGD] (detailing the expenses associated with mock juries).

44. *See infra* Part II.A.

45. *See infra* Part II.B.

46. *See infra* Part II.C.

47. *See infra* Part II.D. The second, third, and fourth steps of this process have been incorporated into the SATE Package's freely available, open-source analysis functions. *See SATE*, *supra* note 12. The researcher simply supplies estimates from the step-one survey to analysis functions to obtain verdict probabilities, margins of error, and graphs of results. Step one requires case-specific analysis while steps two, three, and four are general procedures for making inferences about the probability of a different result based on case-specific data.

48. *See Scientific Framework*, *supra* note 13, at 34.

49. Jurors' initial verdict preferences are based primarily on the strength of evidence against the defendant, but the list of factors thought to influence jurors is extensive. This Article does not discuss how jurors form verdict preferences, but other works analyze jurors' preference formation in detail. Scholars have described the juror's thought process during trial as Bayesian, algebraic, stochastic, story-based, and social cognitive. *See* DENNIS J. DEVINE, JURY DECISION MAKING 22–30 (2012). This analysis does not commit the researcher to identifying a mental model of how jurors make decisions; one simply measures their preferences and calculates the probability of jury verdicts in different trial conditions.

verdict preferences.[50] Once jurors' initial preferences are taken into account, trial-related variables that shape preferences no longer have a direct causal connection to verdicts.[51]

Estimating P(g) is a specific application of public opinion research, but the methods and tools used to identify public sentiment are familiar to lawyers, social scientists, and courts. Litigators use public opinion research methods, including surveys, mock juries, and focus groups to evaluate cases, refine trial strategies, and select jurors.[52] Social scientists routinely administer surveys to estimate quantities like the proportion of adults who favor stronger gun control, the president's approval rating, or the unemployment rate.[53] Surveys are routinely used as evidence in litigation to substantiate factors such as consumer confusion, defamation injuries, and representativeness of a class action and to estimate injuries in groups too large to directly examine.[54] Survey data, such as

50. Jurors' verdict preferences are a link in the causal chain from trials to juror preferences to jury verdicts. Variables like trial evidence, defendant and victim characteristics, the jurisdiction, and jury instructions ultimately affect P(G) because they influence jurors' initial verdict preferences. *See id.* at 55, 93, 122. There are, of course, other variables that affect trials, but they need not concern us here.

51. Once we estimate jurors' initial verdict preferences, we effectively control for all the variables that affect preferences. *See* JUDEA PEARL & DANA MACKENZIE, THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT 113–14 (2018) (describing how a mediator variable in a chain can account for the effects of another variable). If we add variables that explain jurors' initial verdict preferences to an empirical analysis that uses jurors' initial verdict preferences to predict jury verdicts, we will obtain flawed results. *See id.*

52. Focus groups and mock juries are widely used to help litigators prepare for trials. *See* Martha Neil, *Practice Makes Perfect: Mock Trials Gain Ground as a Way to Get Inside Track in Real Trial*, 89 A.B.A. J. 34, 34 (2003); Lisa Spano, Takir Daftary-Kapur & Steven D. Penrod, *Trial Consulting in High-Publicity Cases*, *in* RESEARCH METHODS IN FORENSIC PSYCHOLOGY 215, 216–18 (Barry Rosenfeld & Steven D. Penrod eds., 2011). Lawyers use scientific jury analysis to help pick jurors. *See* Richard Seltzer, *Scientific Jury Selection: Does It Work?*, 36 J. APPLIED SOC. PSYCH. 2417 (2006).

53. Researchers have administered surveys to ascertain the public's opinion on various topics since the 1950s. Public opinion research was first used to study U.S. presidential elections. *See* James W. Tankard Jr., *Public Opinion Polling by Newspapers in the Presidential Election Campaign of 1824*, 49 JOURNALISM & MASS COMMC'N Q. 361, 361 (1972). Early public opinion researchers convincingly demonstrated that relatively small random samples offer better insights into the electorate than very large non-random studies. *See* Peverill Squire, *Why the 1936 Literary Digest Poll Failed*, 52 PUB. OP. Q. 125, 126 (1988). Since those early days of survey research, public opinion research has become standard practice in many fields, especially in politics and business. *See Who We Are*, AM. ASS'N PUB. OP. RSCH., https://aapor.org/about-us/who-we-are/ [https://perma.cc/QN2V-PXAS]. In the legal field, public opinion research comes into play when lawyers use focus groups or surveys to inform jury selection. Surveys also provide evidence in litigation. For example, surveys can be used in trademark litigation to determine whether the public is aware of a brand. *See* Katie Brown, Natasha T. Brison & Paul J. Bautista, *An Empirical Examination of Consumer Survey Use in Trademark Litigation*, 39 LOY. L.A. ENT. L. REV. 237, 239–40 (2019).

54. Courts initially viewed surveys with great suspicion because survey respondents are not subject to cross-examination, but the modern view, initially articulated by Judge Feinberg in *Zippo Mfg. Co. v. Rogers Imps., Inc.*, 216 F. Supp. 670, 682–86 (S.D.N.Y. 1963), is that surveys are, as a general rule, admissible evidence. *See* Schering Corp. v. Pfizer, Inc., 189 F.3d 218, 225 (2d Cir. 1999). In trademark litigation, courts routinely allow expert testimony regarding consumer confusion based on properly administered surveys. *See Schering Corp.*, 189 F.3d at 225. Surveys have also been used to estimate damages suffered by a large class. *See In re* Shell Oil Refinery, 136 F.R.D. 588, 591–92 (E.D. La. 1991).

Although surveys are generally admissible in litigation, they cannot be used for all purposes, and courts have expressed skepticism about some claims based on survey data. *See, e.g.*, Jack Daniel's Props., Inc. v. VIP Prods. LLC, 599 U.S. 140, 163–65 (2023) (Sotomayor, J., concurring) (arguing that surveys may not be reliable due to confusion in trademark infringement cases with constitutionally-protected parodies of

demographic estimates from the U.S. Census Bureau's various survey projects, are also regularly used by courts to document general social conditions in evaluating claims of discrimination.[55]

Professor Rita Simon conducted one of the earliest mock jury experiments as part of the University of Chicago Law School's landmark *American Jury Project*.[56] Although Simon's study is over fifty years old, it remains excellent research on a timeless subject. Simon examined how jurors think about mental illness, criminal responsibility, and the insanity defense.[57] Simon's subjects were real jurors summoned to jury duty in cooperating courts.[58] For part of her research, Simon replicated the *United States v. King* trial,[59] randomly assigning jurors to hear varying definitions of insanity.[60] Simon reported that 76.0% of jurors given the *M'Naghten* rule ($n = 240$) would find the defendant guilty, while 64.0% of jurors given the *Durham* rule ($n = 312$) supported a guilty verdict.[61]

Estimating P(g) is a specialized application of familiar research tools. Many studies analyze survey data to identify factors that generally affect juror verdict preferences, but relatively few studies estimate P(g) values in the context of specific criminal trials like Simon's study of *King*. Table 1 summarizes the results

---

patented products); U.S. Pat. & Trademark Off. v. Booking.com B.V., 591 U.S. 549, 572–74 (2020) (Breyer, J., dissenting) (emphasizing the difficulty in measuring acquired meaning of generic terms using surveys); Atkins v. Virginia, 536 U.S. 304, 326–28 (2002) (Rehnquist, C.J., dissenting) (criticizing the use of public opinion surveys on capital punishment of mentally-impaired individuals).

55. *See, e.g.*, Berghuis v. Smith, 559 U.S. 314, 322–23 (2010) (comparing local jury pools to Census data); Nev. Dep't of Hum. Res. v. Hibbs, 538 U.S. 721, 730 (2003) (citing Bureau of Labor Statistics survey as evidence of need for the Family Medical Leave Act); Zelman v. Simmons-Harris, 536 U.S. 639, 671 (2002) (O'Connor, J., concurring) (citing a Department of Education survey on private school participation in voucher programs).

56. *See* RITA JAMES SIMON, THE JURY AND THE DEFENSE OF INSANITY vii-ix (1967).

57. *Id.* at 12–13.

58. *Id.* at 36.

59. Simon selected *United States v. King*, Dkt. No. 655-5 (D.C. Cir. 1956) as a model to examine the effect of legal definitions of insanity in a non-murder trial. King was prosecuted for incest. SIMON, *supra* note 56, at 50–51. The case was first heard by a jury and later retried before Judge Tamm of the U.S. District Court for the District of Columbia. *Id.* at 50. It is not a published opinion.

Simon also conducted an experiment based on the more famous *Durham* case. *See* Durham v. United States, 214 F.2d 862 (D.C. Cir. 1954). Simon's *Durham* results were not particularly interesting because a substantial majority of jurors would find Durham not guilty by reason of insanity no matter which definition was used. SIMON, *supra* note 56, at 69–70. Following the *M'Naghten* rule, 41% of jurors ($n = 119$) would find the defendant guilty compared to 35% of jurors ($n = 119$) following the *Durham* instruction. SIMON, *supra* note 56, at 68 tbl. 8. In both conditions, the probability of a guilty verdict is low (.190 compared to .110). SIMON, *supra* note 56, at 68 tbl. 8. Monte Durham was, in fact, found not guilty by reason of insanity. Durham, 214 F.2d at 865. There is little reason to assess the harmfulness of the jury instruction when the defendant was found not guilty.

60. *See Durham*, 214 F.2d at 874–75; M'Naghten's Case (1843) 8 Eng. Rep. 718, 719–20. For background and discussion of the contrasting definitions of insanity that emerge from these cases, see Simon E. Sobeloff, *Insanity and the Criminal Law: From McNaghten to Durham, and Beyond*, 41 A.B.A. J. 793, 793–96 (1955); Abe Krash, *The Durham Rule and Judicial Administration of the Insanity Defense in the District of Columbia*, 70 YALE L.J. 905, 921–38 (1960).

61. SIMON, *supra* note 56, at 72 tbl. 10. Simon reported results to two decimal places (e.g., .76 and .64). I am using .760 and .640 for consistency with other studies.

of studies that estimate verdict preferences in jury pools for real criminal trials.[62] The author conducted the *Porter*, *Washington*, *Fulminante*, and *Hopkins* studies.[63] Christopher Robertson and his students studied the *Chapman* and *Lee* trials.[64] Dan Kahan studied the impact of jury instructions in the *Berkowitz* case.[65]

Table 1. Verdict Preferences Estimated in Real Criminal Trials

Trial	Error/Issue	P(g | actual)	P(g | hypo)
U.S. v. King	Jury instructions	76.0% (240)	64.0% (312)
Fla. v. Porter	Ineffective counsel	77.3% (761)	60.5% (761)
Fla. v. Washington	Ineffective counsel	87.9% (773)	91.1% (773)
Ariz. v. Fulminante	Coerced confession	46.3% (764)	41.4% (764)
Tex. v. Hopkins	Coerced confession	85.8% (917)	88.3% (917)
Cal. v. Chapman	Prosecutor misconduct	60.0% (99)	53.0% (99)
Pa. v. Lee	Prosecutor misconduct	68.0% (69.5)	66.0% (69.5)
Pa. v. Berkowitz	Jury instructions	65.0% (300)	53.0% (300)

Note: Sample sizes reported in parentheses.

In these studies, respondents are randomly assigned to either the actual trial condition or a hypothetical trial condition.[66] Subjects randomly assigned to the actual trial condition read a summary of the evidence that was presented at the defendant's trial while subjects randomly assigned to the hypothetical trial condition read a summary of the evidence that would be presented at an error-

62. These studies are discussed further *infra* Part III.

63. Replication files and survey datasets are available online. *See* Barry C. Edwards, *Replication Data for: Measuring Fairness*, HARVARD DATAVERSE (Jan. 21, 2026), https://doi.org/10.7910/DVN/RPQT5X (online repository for *Porter*, *Washington*, *Fulminante*, and *Hopkins* survey data and replication code) [hereinafter Edwards, *Survey Results*].

64. *See* D. Alex Winkelman, et al., *An Empirical Method for Harmless Error*, 46 ARIZ. ST. L.J. 1405, 1418–21 (2014). The article reports a total sample size of 198 for the *Chapman* study but does not specify the sample size in each condition. For present purposes, I assume $n = 99$ in each condition. For the *Lee* study, I assume $n = 69.5$ in each condition, based on the overall reported sample size of 139. The samples are not weighted to represent specific jurisdictions, nor are they limited to respondents who meet standard jury qualifications. We are limited to analyzing the reported verdict preferences.

65. *See* Dan M. Kahan, *Culture, Cognition, and Consent: Who Perceives What, and Why, in Acquaintance-Rape Cases*, 158 U. PA. L. REV. 729, 765 (2010). Kahan's experiment varied the legal standards that subjects (n = 1,500, a nationally representative sample, not restricted to jury-qualified respondents) were asked to apply. *Id.* at 779 tbl. 1. There were five experimental conditions. *Id.* at 767–69. For purposes of calculating standard errors, I assume $n = 300$ in each condition.

66. The jury instruction studies are not examples of trial errors. Simon and Kahan examined the effect of jury instructions in the context of real criminal trials. *See* SIMON, *supra* note 56, at 50; Kahan, *supra* note 65, at 796–97. The author of this Article designated actual and hypothetical trial conditions in the *King* and *Berkowitz* studies.

free trial.[67] Respondents are not told which version they are reading or that there was any error in the trial.[68]

After respondents are randomly assigned to read a summary of the actual or hypothetical trial condition, the author's surveys instruct them to "[c]onsider evidence that has been presented to you about this case" and then ask whether they support a guilty or not guilty verdict (or life imprisonment or the death penalty).[69] To maximize the amount of data collected from a sample, the author

67. These surveys should accurately estimate verdict preferences in the population that courts have in mind when assessing the probability of a different jury verdict. Surveys, like juries, are meant to convey the views of a cross-section of a community. The trial summaries used in surveys do not provide respondents a virtual-reality-type experience of the trial, but they do generate valid verdict preferences. The validity of surveys in this context is an important consideration. Researchers have demonstrated the validity of estimating jurors' verdict preferences using survey respondents' responses to written vignettes. *See, e.g.*, Steffen Bieneck, *How Adequate is the Vignette Technique as a Research Tool for Psycho-Legal Research*, *in* SOCIAL PSYCHOLOGY OF PUNISHMENT OF CRIME 255, 263–64 (Margit E. Oswald, Steffen Bieneck & Jörg Hupfeld-Heinemann eds., 2009); Michael J. Saks, *What Do Jury Experiments Tell Us About How Juries (Should) Make Decisions?*, 6 S. CAL. INTERDISC. L.J. 1, 9–11 (1997). Researchers have similarly shown that surveys conducted online generate valid results. *See, e.g.*, Alexander Coppock, *Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach*, 7 POL. SCI. RSCH. & METHODS 613, 613–14 (2019); Christoph Bartneck, et al., *Comparing the Similarity of Responses Received from Studies in Amazon's Mechanical Turk to Studies Conducted Online and with Direct Recruitment*, 10(4) PLOS ONE (2015) [https://perma.cc/DA49-RDJX].

While comparative evaluations of different methods of measuring preferences are generally supportive, the researcher should purposefully design studies to accurately estimate verdict preferences in the relevant population. *See* Barry Edwards, *If the Jury Only Knew: The Effect of Omitted Mitigation Evidence on the Probability of a Death Sentence*, 32 VA. J. SOC. POL. & L. 1, 18–33 (2025) (discussing litigation context and the representativeness of subject pools).

68. Edwards, *Survey Results*, *supra* note 63; Winkelman et al., *supra* note 64; Kahan, *supra* note 65.

69. The answer order is varied to neutralize answer-order effects. This question offers only two alternatives and forces respondents to choose one of them. There is no "Unsure," "Don't know," or "Prefer not to say" alternative. *See* Edwards, *Survey Results*, *supra* note 63; Winkelman et al., *supra* note 64; Kahan, *supra* note 65. There is some debate over forcing survey respondents to choose answers without offering a "Don't know" (DK) response, but the option is not advisable when respondents can make a choice. *See generally* Jon A. Krosnick et al., *The Impact of "No Opinion" Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice?*, 66 PUB. OP. Q. 371 (2002) (suggesting that including no-opinion options precludes data measurement of meaningful opinions). One review of the literature concludes that

> [DKs] often result not from genuine lack of opinion but rather from ambivalence, question ambiguity, satisficing, intimidation, and self-protection. In each of these cases, there is something meaningful to be learned from pressing respondents to report their opinions. [But DK] options discourage people from doing so under these circumstances. This explains why data quality is not improved when such options are explicitly included in questions.

Jon A. Krosnick, *The Causes of No-Opinion Responses to Attitude Measures in Surveys: They Are Rarely What They Appear to Be*, *in* SURVEY NONRESPONSE 87, 99 (R.M. Groves et al. eds., 2002); *see also* Patrick Sturgis, Caroline Roberts & Patten Smith, *Middle Alternatives Revisited: How the Neither/nor Response Acts as a Way of Saying "I Don't Know"*, 43 SOCIO. METHODS & RSCH. 15, 17–18 (2014) (finding that, in the majority of cases, "don't know" responses significantly decrease the accuracy of surveys); Daniela Wetzelhütter, *Scale-Sensitive Response Behavior!? Consequences of Offering Versus Omitting a "Don't Know" Option and/or a Middle Category,* 13 SURV. PRAC., https://www.surveypractice.org/article/14484-scale-sensitive-response-behavior-consequences-of-offering-versus-omitting-a-don-t-know-option-and-or-a-middle-category [https://perma.cc/EDC5-KC3S] (discussing advantages of using middle categories and "don't know" options and characterizing selection of these as "most probably a compromise"); Daniel Laurison, *The Willingness to State an Opinion: Inequality, Don't Know Responses, and Political Participation*, 30 SOCIO. F. 925, 944 (2015) (demonstrating that "don't know" responses in surveys are stratified by income, skewing accuracy due to self-perceptions of lower-income respondents).

uses a crossover design.[70] After describing the differences in the new trial condition, subjects are again asked to express their verdict preference.[71] The crossover design thereby balances the experimental groups.[72] The design is also necessary to evaluate how a trial error affects jurors' perceptions and understandings, apart from their verdict choices.[73]

The author recruited survey respondents from an online platform. Because online survey respondents are not fully representative of the adult population in a court jurisdiction, in the author's studies, responses are weighted to represent the relevant jurisdiction.[74] Additionally, P(g) estimates are restricted to respondents who meet standard jury qualifications as well as qualifications for death penalty cases. Despite these restrictions, the author's sample sizes are relatively large compared to other studies of real criminal trials.[75]

While varied designs and research protocols are appropriate for general research, there need to be standard procedures for estimating P(g) in different trial conditions to calculate the probability of a different outcome. The studies summarized in Table 1 use different subject pools, presentation materials, and

---

70. Medical researchers often use crossover research designs. Subjects who were randomly assigned to the control group are later switched to the treatment group, and those in the treatment group are switched to the control group. *See* Thomas A. Louis et al., *Crossover and Self-Controlled Designs in Clinical Research*, 310 NEJM 24, 24 (1984). The outcome of interest is measured at the end of the first phase and again at the end of the second phase. *Id.*

71. *See* Edwards, *Survey Results*, *supra* note 63; Winkelman et al., *supra* note 64; Kahan, *supra* note 65.

Doubling the sample size used to estimate P(g | actual) and P(g | hypothetical) reduces the margin of error of these estimates by the square root of 2. *See* Hamed Taherdoost, *Determining Sample Size; How to Calculate Survey Sample Size*, **2** INT'L J. ECON. & MGMT. SYS. 237, 237 (2017). Efficiency is desirable, but the cost of recruiting participants should not dictate the research design or choice between approaches to harmless error analysis. It is a factor to consider, of course, but it is not as important as accurate estimates and analysis.

Some might ask whether historical evidence from actual trials could provide some additional data based on the six or twelve actual jurors who ultimately returned a guilty verdict. While the historical record could add six or twelve additional data points, those are relatively small samples compared to sample sizes that can be obtained at reasonable cost today. Moreover, while the actual verdict is known, the initial opinions of jurors prior to deliberation may not be a matter of record.

72. *See* Louis et al., *supra* note 70, at 24–25. If the researcher does not use a crossover design, is there a good solution to the possible non-compliance issues of subjects disproportionately quitting the longer summaries? *See* Adeline Lo, Jonathan Renshon & Lotem Bassan-Nygate, *A Practical Guide to Dealing with Attrition in Political Science Experiments*, 11 J. EXPERIMENTAL POL. SCI. 147, 147–48 (2024) (explaining that attrition in studies can affect validity of results).

73. One approach to harmless error analysis does not view the jury as simply a means of deciding trial outcomes, but rather a potential victim of trial errors. *See* Edwards, *Scientific Framework*, *supra* note 13, at 19 ("effect-on-trial approach"). In this view, important trial outcomes include establishing truths and community values. *Id.* at 17. I discuss this view as the effect-on-jury approach. To evaluate the effect-on-jury, it is necessary to describe the difference between trial conditions and ask respondents how much it affects their thoughts (even if it does not change their vote). *See id.* at 45.

74. *See* Edwards, *Survey Results*, *supra* note 63. Because the population that fills out online surveys is not demographically representative of the total population, weighing online survey responses can alleviate the selection bias of the sample actually gathered. *See* Chadia Haddad et al., *Should Samples be Weighted to Decrease Selection Bias in Online Surveys During the COVID-19 Pandemic? Data from Seven Datasets*, 22 BMC MED. RSCH. METHODOLOGY, at 2 (2022).

75. *See, e.g.*, Dennis J. Devine et al., *Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups*, 7 PSYCH., PUB. POL'Y, & L. 622, 628 tbl.1, 29 tbl.1 (2001).

analysis techniques.[76] The researcher estimating P(g) for a specific trial should think carefully about the population the sample is meant to represent, how trial conditions are presented to respondents, survey instructions and questions, and weighting observations for analysis.[77] There are not yet standard protocols and procedures for estimating P(g) values, but the researcher may commit to a specific research design and plan for analysis before collecting and analyzing data to prevent changing procedures and produce a favorable result.[78] The researcher should save all materials necessary to verify the analysis and independently replicate the results.[79] The researcher should aim only to produce accurate estimates, without favoring one side or the other.

B. Initial Verdict Factions Given Preferences of Jury Pool

Following the process discussed in the last Part, P(g) estimates verdict preferences in a jury pool from which trial jurors are selected.[80] Simon, for

76. *See* Edwards, *Survey Results*, *supra* note 63; Winkelman et al., *supra* note 64; Kahan, *supra* note 65.

77. Additional considerations for the respondent recruitment plan include target sample size, task description, compensation, and recruitment platform.

78. In social sciences, committing to a design and plan for analysis is known as pre-registration. *See* Macartan Humphreys, Raul Sanchez de la Sierra & Peter van der Windt, *Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration*, 21 POL. ANALYSIS 1, 1–2 (2013); Garret Christensen & Edward Miguel, *Transparency, Reproducibility, and the Credibility of Economics Research*, 56 J. ECON. LITERATURE 920, 955 (2018). It is standard practice for the researcher to pilot test a survey to make sure that it is working properly. *See, e.g.*, Winkelman et al., *supra* note 64, at 1421; Krosnick et al., *supra* note 69, at 382–83. The researcher might, for example, pilot test a survey to verify proper randomization, check completion times, and make sure responses are recorded properly. Responses obtained from pilot testing are not analyzed to estimate P(g).

Collaborating with opposing parties and committing to research design before collecting data may help guard against the human tendency to judge a research process based on whether its results confirm prior beliefs. This type of reasoning is evident in cases addressing survey methods. According to one court: "[A] skeptic would classify the survey cases into two categories: a survey is accepted and relied upon when the judge already has his or her mind made up in favor of the survey results; and a survey is rejected and torn apart when the judge subjectively disagrees with the survey results." Schering Corp. v. Pfizer Inc., 189 F.3d 218, 227 (2d Cir. 1999) (quoting J. THOMAS MCCARTHY, 5 MCCARTHY ON TRADEMARKS AND UNFAIR COMPETITION § 32:196 (4th ed. 1998)).

79. The researcher should make individual-level data available along with the computer code used to analyze the data. Individual-level data should give others the opportunity to verify that the subject pool has not been manipulated. For example, raw survey data should include IP addresses of subjects to verify that "ballot box stuffing" from a single IP address has not occurred. If the method is applied to ongoing cases, the researcher should make sure interested parties do not actively participate in the research by responding to online surveys or directing others to do so. The researcher can, for example, keep the start time secret and recruit participants from different forums. Additionally, the researcher should save recruitment data; Amazon's MTurk compiles the identification numbers of those who completed a task through its system. *See Mechanical Turk (MTurk)*, OFF. OF RSCH. ETHICS, IOWA ST. UNIV. 3 (June 2, 2020), https://compliance.iastate.edu/wp-content/uploads/sites/4/pdf/MTurk-guidance.pdf [https://perma.cc/7V65-FKNM]. Neither MTurk identification numbers nor IP addresses recorded in the survey data identify subjects by name for privacy reasons, but they would allow others to verify that data have not been collected inappropriately. *See id.* at 2.

80. See *infra* Parts II.B and III for examples of studies that estimate the proportion of jurors who would prefer a guilty verdict under varying trial conditions. These studies fairly estimate the verdict sentiments

example, found that 76.0% of jurors would convict defendant King under the *M'Naghten* rule and 64.0% of jurors would convict given the *Durham* rule.[81] Estimating P(g) is a necessary start, but it is distinct from the probability of a guilty verdict, P(G), which is the real quantity of interest.

1. Initial Verdict Factions with Random Selection

Given the value of P(g) in the jury pool, one can calculate the probabilities of selecting $k = 0, 1, 2, 3 \ldots n$ jurors who would support a guilty verdict using a venerable statistical rule.[82] If the jury pool is large compared to the size of the jury, the probability of randomly selecting k jurors who would prefer a guilty verdict on a jury with n members follows a binomial distribution, a generally accepted principle used by courts to evaluate probabilities.[83]

$$P(k \mid P(g)) = \frac{n!}{k!\,(n-k)!}(P(g))^{k}(1-P(g))^{n-k}$$

$P(k \mid P(g))$ represents the probability of selecting k jurors who will vote for a guilty verdict when P(g) is the proportion of the jury pool that prefers a guilty verdict.[84] To illustrate, Figure 1 depicts expected initial guilty votes when twelve-person juries are randomly drawn from jury pools where P(g) = .760 and P(g) = .640, the verdict preferences in *King* estimated by Simon based on the *M'Naghten* and *Durham* rules.[85]

---

of a large jury pool, but one would not observe the same proportion of guilty votes on a trial jury due to random selection of jurors from the pool and the possible use of peremptory strikes.

81. SIMON, *supra* note 56, at 72 tbl. 10.

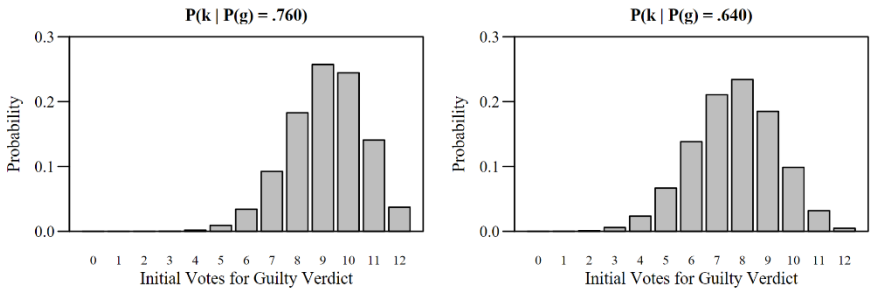
82. *See* ALAN AGRESTI & BARBARA FINLAY, STATISTICAL METHODS FOR THE SOCIAL SCIENCES 169–72 (4th ed. 2009) (explaining binomial distribution); KOSUKE IMAI, QUANTITATIVE SOCIAL SCIENCE: AN INTRODUCTION 282–86 (2018) (same); *see also* Jaime Israel García-García et al., *The Binomial Distribution: Historical Origin and Evolution of Its Problem Situations*, 10 MATHEMATICS, 6–28 (2022) (discussing history of the binomial distribution).

83. *See* Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1970) (discussing binomial probability distribution and citing authoritative references); *see also* Krim v. pcOrder.com, Inc., 402 F.3d 489, 492–93 n.6 (5th Cir. 2005) (discussing statistical evidence based on "elementary principles of binomial probability"); Washington v. People, 186 P.3d 594, 603 (Colo. 2008) (en banc) (using binomial probability distribution to evaluate jury selection practices); Moultrie v. Martin, 690 F.2d 1078, 1083 n.8 (4th Cir. 1982) (citing and discussing the binomial distribution formula used to evaluate grand jury selection practices).

84. It is not necessary or advisable to calculate $P(k \mid P(g))$ probabilities manually as it is a standard feature of data analysis software. *See* Clemma J. Muller & Richard F. MacLehose, *Estimating Predicted Probabilities from Logistic Regression: Different Methods Correspond to Different Target Populations*, 43 INT'L J. EPIDEMIOLOGY 962, 963 (2014) (explaining how data analysis programs like SAS and Stata often default to calculating conditional probability). This may look like a thorny equation, but it simplifies considerably when we insert specific numerical values for P(g), k, and n into the equation.

85. SIMON, *supra* note 56, at 72 tbl. 10.

Figure 1. Probabilities of Initial Guilty Vote Counts Given Jury Pool Preferences



On average, one would expect twelve-person juries picked from these pools to have 9.12 and 7.68 initial votes for guilty verdicts,[86] but it is impossible to pick those specific values, and a variety of starting polls are possible. The distributions illustrated in Figure 1 assume jurors are randomly selected from the jury pool. This starting assumption oversimplifies the jury selection process, but it is possible to model the probability of a guilty verdict following the strategic use of peremptory strikes.

2. *Accounting for Peremptory Strikes*

To more accurately represent the selection of jurors from jury pools, P(*k* | P(g)) values can be calculated in a way that accounts for parties using peremptory strikes. Peremptory strikes can be addressed by specifying how many strikes are used by the prosecution and defense, respectively. These are parameters, like jury size, that are specified by the jurisdiction before deliberation begins.

Peremptory challenges do not change juror preferences, but they do modify the distribution of preferences observed on trial juries.[87] When parties accurately strike potential jurors, with the prosecution striking not guilty votes and the defense striking guilty votes, fewer twelve-person juries start with seven

86. The average number of votes can be calculated by taking the expectation of the binomial distribution—multiplying each result (number of guilty votes) by its respective probability.

87. Rather than selecting just enough jurors from the pool to fill the jury box, extra jurors are selected to account for peremptory challenges. For example, in a federal non-capital felony trial, the prosecution is entitled to six peremptory challenges, and the defendant is allowed to strike ten jurors. FED. R. CRIM. P. 24(b)(2). To model this process, twenty-eight potential jurors are randomly selected from the jury pool. The parties' combined sixteen strikes leave twelve jurors for trial. If twenty-two or more of the twenty-eight potential jurors favor a guilty verdict, the twelve who remain after parties use their challenges will all support a guilty verdict; the defendant does not have enough strikes to keep them all off the jury. If ten or fewer jurors out of the twenty-eight potential jurors favor a guilty verdict, there will be no votes for a guilty verdict among the final twelve, as the defendant struck them all.

or eight votes for a guilty verdict and more juries start with strong majorities for one verdict or the other.[88]

The net effect of peremptory challenges is mild for several reasons. Both sides exercise challenges, and they tend to cancel each other out. Additionally, parties cannot accurately identify and strike jurors likely to vote against them. The characteristics that litigators can observe during voir dire are weak predictors of how jurors will vote.[89] Research indicates that variables attorneys could obtain from juror questionnaires explain a modest amount of variation in jurors' initial verdict preferences, about ten to fifteen percent.[90] In reality, peremptory strikes are closer to a random process than a procedure executed with perfect accuracy.[91] The default setting in analysis routines assumes parties strike jurors imperfectly.[92]

Although peremptory strikes play a limited role, there is no reason to ignore them. In most cases, peremptory strikes can be addressed by simply specifying the number of strikes available to the prosecution and the defense. For example, the prosecution and defense would each be entitled to three strikes in *King*, a felony trial in the District of Columbia.[93] While peremptory strikes do not dramatically change fairness estimates, we may account for them to measure fairness as accurately as possible.

C. Probability of Guilty Verdict Based on Initial Guilty-Verdict Faction

This Part outlines the jury deliberation model that generates the probability of a guilty verdict based on the number of jurors who initially favor a guilty verdict. Models are simplified representations of complex systems; a useful

88. Modifying the original twelve-juror distribution by adding jurors to be struck spreads the binomial distribution thinner, leading to a lower probability of the jury being evenly divided when the prosecution and defense accurately strike those who would vote against their clients if seated on the jury.

89. It is hard to predict how potential jurors will vote based on characteristics that attorneys could observe before trial. Contrary to common belief, for example, social science research indicates that female jurors are generally not more sympathetic to criminal defendants or civil plaintiffs. *See* Dennis J. Devine et al., *Strength of Evidence, Extraevidentiary Influence, and the Liberation Hypothesis: Data from the Field*, 33 L. & HUM. BEHAV. 136, 142 (2009); Solomon M. Fulero & Steven D. Penrod, *Attorney Jury Selection Folklore: What Do They Think and How Can Psychologists Help?*, 3 FORENSIC REPORTS 233, 236–37 (1990). *But see* John K. Cochran & Beth A. Sanders, *The Gender Gap in Death Penalty Support: An Exploratory Study*, 37 J. CRIM. JUST. 525, 530 (2009) (discussing why women support the death penalty less than men).

90. *See* DEVINE, *supra* note 49, at 54; Seltzer, *supra* note 52, at 2422–23 (discussing modest explanatory power reported in studies of scientific jury selection); Joel D. Lieberman, *The Utility of Scientific Jury Selection: Still Murky After 30 Years*, 20 CURRENT DIRECTIONS IN PSYCH. SCI. 48, 49–50 (2011).

91. *See* Lieberman, *supra* note 90, at 49–50. If the parties strike jurors at random, the expected distribution of votes after strikes would be the same as if no strikes were allowed.

92. The SATE statistical software package that implements this analysis uses 15% peremptory strike accuracy as its default setting. *See Scientific Analysis of Trial Errors (SATE)*, R PROJECT (Nov. 5, 2025), https://cran.r-project.org/web/packages/sate/sate.pdf [https://perma.cc/MA7B-WWU6]. This means that parties use 15% of their strikes to remove jurors who would vote against them; their other strikes are applied at random.

93. *See* D.C. CODE § 23-105(a).

model omits unnecessary details but retains the essential features of the system represented.[94] Here, the focus is solely on the jury's verdict, without modeling other aspects of jury deliberation, such as the selection of the foreperson, the frequency or content of jurors' discussions, or the duration of deliberation.

This Part begins by identifying the essential features of the jury-deliberation process that are represented by a mathematical construct called a Markov chain.[95] This model generates the probability that deliberation will end with a guilty verdict given the jury's initial state. The model is validated by comparing its predictions to verdicts observed in five decades of jury deliberation research. The Markov chain model of deliberation presented here provides more precise and reliable estimates of verdict probability than is possible through empirical analysis alone. In this Part, I describe deliberation as a Markov chain process, calculate the model's guilty-verdict probabilities, and compare model predictions to observed deliberation results. This synthetic model distills the essential features of jury deliberation and produces results consistent with empirical analysis.

1. Jury Deliberation as a Markov Chain with Absorbing States

Most juries take a preliminary vote to reveal how many jurors support a guilty verdict (represented as k) and how many support a not guilty verdict ($n - k$). They are not instructed to take a vote; they simply vote spontaneously by raising their hands, counting ballots, or by announcing their choice between guilty and not guilty verdicts one at a time.[96] Figure 2 depicts a jury with $k = 6$. The value of k is a function of P(g) and the jury size (n).[97]

---

94. "All models are wrong but some are useful." This well-known statement was first used as a subheading in a paper by George Box. *See* George E. P. Box, *Robustness in the Strategy of Scientific Model Building*, *in* ROBUSTNESS IN STATISTICS 201, 202 (Robert L. Launer & Graham N. Wilkinson eds., 1979) (cleaned up).

95. The Markov chain model is discussed and illustrated with figures *infra* Part II.C.2. The Markov chain construct and technique of simulating model outcomes to obtain results (known as Markov Chain Monte Carlo, or MCMC) is a widely used and generally accepted method of analysis. *See, e.g.*, United States v. Lewis, 442 F. Supp. 3d 1122, 1141 (D. Minn. 2020) ("[W]ell-known mathematical modeling process that is used to solve complex problems across a wide array of disciplines including weather forecasting, physics, and engineering."); People v. Wakefield, 195 N.E.3d 19, 24 (N.Y. 2022) (noting that the method is "well-established"); Whittley v. State, 2022 WL 3645589, No. 417-81486-2020, slip op. at *15 (Tex. App. Aug. 24, 2022) (describing MCMC as "a widely accepted mathematical modeling technique used in many disciplines, such as code breaking, physics, social science, and computer linguistics"); People v. Davis, 290 Cal. Rptr. 3d 661, 667 (Cal. Ct. App. 2022) (noting MCMC is among "well-established and accepted mathematical principles"). *See also* Harkenrider v. Hochul, 197 N.E.3d 437, 460 (N.Y. 2022) (Wilson, J., dissenting) (noting the method is "regularly used in redistricting cases"). The Markov chain is named after Russian mathematician Andrey Markov. *See* Gely P. Basharin, Amy N. Langville & Valeriy A. Naumov, *The Life and Work of A.A. Markov*, 386 LINEAR ALGEBRA & ITS APPLICATIONS 3, 3 (2004).

96. Juries typically take their first poll twenty to forty-five minutes into deliberations, before any substantial discussion of the trial. *See* Shari Seidman Diamond et al., *Juror Discussions During Civil Trials: Studying an Arizona Innovation*, 45 ARIZ. L. REV. 1, 61 (2003); *see also* Valerie P. Hans et al., *The Hung Jury: The American Jury's Insights and Contemporary Understanding*, 39 CRIM. L. BULL. 33, 37 (2003) (discussing implications of initial jury poll).

97. *See infra* Part II.B.

The jury must deliberate until all jurors align with one side or the other. If the jury's initial poll is unanimous ($k = 0$ or $k = n$), there is no need to persuade anyone to change their vote.[98] If the jury's initial vote is not unanimous, the guilty-verdict faction deliberates with the not-guilty-verdict faction, each attempting to pull members of the opposing faction to its side until everyone is on the same side.

While the jury deliberates, one member of each faction is closest to joining the opposing faction and must decide whether to stay with their faction or switch sides.[99] In this model, p_k represents the probability that persuadable jurors will join the guilty-verdict faction when there are currently k votes for a guilty verdict.[100] The specific reasons for their personal ambivalence are too varied to enumerate, but they may be generally summarized as personal factors—jurors have different personalities, life experiences, and thoughts about the trial and defendant. For these jurors, staying with their current faction and switching sides are equally appealing options. This does not imply or require that all jurors' private thoughts are equally balanced between two possible verdicts all the time. It simply means that there are moments in jury deliberation when one member of each faction is on the fence between factions and could switch sides; indeed, for the jury to reach a verdict, k or $n - k$ jurors must eventually switch sides.

---

98. The jurors would still need to elect a foreperson and may discuss the trial, but they do not need to deliberate to create agreement because they already agree on the verdict.

99. *See infra* Figure 2. Two votes, one from each faction, are in play during each stage of deliberation.

100. The quantity p_k represents the probability that a juror on the fence between verdict factions will opt for a guilty verdict in the next stage of deliberation. The formula for p_k is derived from the key features of deliberation. Specifically, p_k combines the external influence of the guilty-verdict faction and the internal, personal factors that make certain jurors susceptible to switching sides:

$$p_k = w_{ex}(external\ factors) + w_{in}(internal\ factors)$$

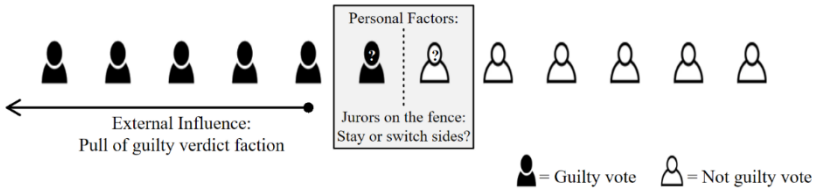
where w_{ex} represents the relative weight of external factors and w_{in} represents the relative weight of internal factors. The relative weight of external and internal factors likely varies among jurors, but they are assumed to have equal weight, on average.

The external factor influencing jurors in deliberation is the current size of the guilty-verdict faction. The influence of the guilty-verdict faction is equal to its relative size, excluding the one member who is susceptible to changing sides. Thus, the force of external factors is equal to $(k - 1) / n$ where k is the current number of votes for a guilty verdict and n is jury size. Operationalizing social influence as $(k - 1) / n$, as opposed to k / n, is based on the burden of proof requirement in criminal cases and consistent empirical evidence of a leniency bias in jury deliberation. *See* Robert J. MacCoun & Norbert L. Kerr, *Asymmetric Influence in Mock Jury Deliberation: Jurors' Bias for Leniency*, 54 J. PERSONALITY & SOC. PSYCH. 21, 30 (1988). Internal factors influencing jurors during deliberation include the personal experiences and private thoughts that shape a juror's views on guilt and innocence. A juror torn between verdict factions, divided between guilty and not guilty, has a one-half probability of siding with the guilty-verdict faction, based on internal factors. Substituting external factors = $(k - 1) / n$, internal factors = ½, w_{ex} = ½, and w_{in} = ½ into the general expression of p_k, above, we obtain the following expression:

$$p_k = \frac{1}{2}\left(\frac{k-1}{n}\right) + \frac{1}{2}\left(\frac{1}{2}\right)$$

which simplifies to the p_k formula presented in the text with Figure 3.

Figure 2. Illustration of Deliberation between Factions



The probability of persuadable jurors switching sides is a combination of their personal preferences, which are divided, and the relative size of the two competing factions. A verdict faction's capacity to pull persuadable jurors to its side is a function of its size and the burden of proof necessary for conviction.[101] The guilty-verdict faction's capacity to pull persuadable jurors to its side is equal to its relative size less its one member who has doubts and may switch sides. When the guilty-verdict faction gains or loses a vote, the probability of its gaining or losing another vote in the next stage of deliberation increases.

There are three possible outcomes in a round of deliberation: the guilty-verdict faction gains one member, loses one member, or stays the same size. The guilty-verdict faction increases by one if both jurors opt for a guilty verdict, which occurs with probability $(p_k)^2$. The guilty-verdict faction decreases by one if both jurors opt for a not guilty verdict, which occurs with probability $(1 - p_k)^2$. The number of guilty votes may also remain unchanged, with probability equal to $2p_k(1 - p_k)$.

The essential features of deliberation are represented as a Markov chain, a mathematical model of a system whose state changes in discrete steps.[102] At each step, the system transitions from one state to another.[103] The future state of the system depends solely on its current state, regardless of how it reached that state.[104] A Markov chain with absorbing states is an extension of the standard Markov chain. Once the system enters an absorbing state, it remains

101. *See* MacCoun & Kerr, *Asymmetric* Influence, *supra* note 100, at 30 (finding acquittal is more favored under the reasonable-doubt standard).

102. *See generally* WAI-KI CHING & MICHAEL K. NG, MARKOV CHAINS: MODELS, ALGORITHMS AND APPLICATIONS (2006) (providing technical background for Markov Chain models); PAUL A. GAGNIUC, MARKOV CHAINS: FROM THEORY TO IMPLEMENTATION AND EXPERIMENTATION (2017) (explaining applications of a Markov chain model generally). Courts have recognized the general acceptance of Markov chain models. *See* cases cited *supra* note 83.

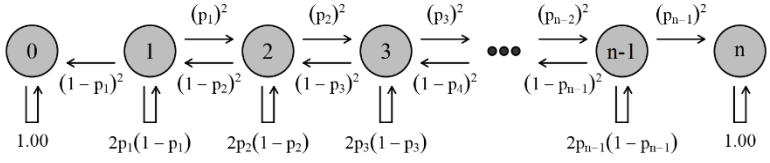
103. Markov chains are widely used in various fields, including physics, economics, biology, and computer science, to model and analyze systems with stochastic (random) behavior. *See, e.g.*, Andrew D. Martin & Kevin M. Quinn, *Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999*, 10 POL. ANALYSIS 134, 135 (2002); Bruce A. Craig & Peter P. Sendi, *Estimation of the Transition Matrix of a Discrete-Time Markov Chain*, 11 HEALTH ECON. 33, 33 (2002); Jess Benhabib, Alberto Bisin & Mi Luo, *Wealth Distribution and Social Mobility in the US: A Quantitative Approach*, 109 AM. ECON. REV. 1623, 1626 (2019).

104. The "memoryless" property is the defining feature of Markov chains. *See* RICHARD DURRETT, ESSENTIALS OF STOCHASTIC PROCESSES 1 (Richard DeVeaux, Stephen E. Fienberg & Ingram Olkin, eds., 3d ed. 2016).

there and cannot transition to any other state.[105] Absorbing states are terminal, representing outcomes from which no further movement occurs.[106]

Applied to jury deliberation, a Markov chain model represents the possible states of the guilty-verdict faction and the probabilities of transitioning between them during deliberation.[107] For a jury with *n* members, the system's possible states are 0, 1, 2 . . . *n*. The initial state of jury deliberation can correspond to any node on the chain, including its endpoints. If the initial state is neither 0 nor *n*, the deliberation process continues until the system reaches one of its endpoints, at which point the system no longer changes its state. Jurors may switch sides slowly, in rapid succession, or perhaps after some back-and-forth. There is no time limit or set number of rounds for deliberation. At some point, the jurors align on one side, the deliberation process is over, and the jury returns its verdict. The Markov Chain model is depicted in general form for *n*-person juries in Figure 3.[108]

Figure 3: Jury Deliberation Process as a Markov Chain



where $p_k = \frac{k-1}{2n} + \frac{1}{4}$

105. *See* JAMES R. NORRIS, MARKOV CHAINS 11 (1997) (explaining that once a Markov chain is in an absorbing state, there is no escaping it).

106. *Id.* One can also imagine the jury's deliberation as the path of a ball on a modified Galton board. A Galton board, introduced to popular culture by The Price is Right's "Plinko" game, is a mechanical device that can be used to show that a well-defined normal distribution is generated by a random process. *See* IMAI, *supra* note 82, at 303–04. In 1877, Francis Galton used the device for a live demonstration of regression to the mean before the Royal Institution of London. *See* PEARL & MACKENZIE, *supra* note 51, at 52–57. A ball traveling down a Galton board has an equal probability of moving left or right when it hits a pin. *Id.* at 54–57. To represent the path of jury deliberation, the probabilities of moving left or right would need to vary across the board as if the pins were set at angles or the board were curved.

107. Abstract models, solved mathematically or with computer simulation, have been used to study jury deliberation, with notable works published in the late 1970s and early 1980s. These models often attempt to model individual participation, deliberation duration, the incidence of hung juries, and additional outcomes relevant to the then-hot debate over ideal jury size. *See, e.g.*, Subhash C. Narula & Vasant Katdare, *Markov Chain Theory and Jury Deliberation Process*, 2 COMPUTS. & OPERATIONS RSCH. 109, 111 (1975); Steven Penrod & Reid Hastie, *A Computer Simulation of Jury Decision Making*, 87 PSYCH. REV. 133, 134 (1980); REID HASTIE, STEVEN D. PENROD & NANCY PENNINGTON, INSIDE THE JURY 175–95 (1983); Alvin K. Klevorick & Michael Rothschild, *A Model of the Jury Decision Process*, 8 J. LEGAL STUD. 141, 146–47 (1979); Bernard Grofman, *Mathematical Models of Juror and Jury Decision-Making: The State of the Art*, *in* THE TRIAL PROCESS 305, 324–30 (Bruce Dennis Sales ed.,1981); Sarah Tanford & Steven Penrod, *Computer Modeling of Influence in the Jury: The Role of the Consistent Juror*, 46 SOC. PSYCH. Q. 200, 201–02 (1983).

108. The transition probabilities in Figure 3 are stated in general form for juries of any size. When *n* is set to a specific value, like 6 or 12, the transition probabilities simplify to real numbers.

Unlike the famous Gambler's Ruin problem, where the probability of transition to a higher (or lower) state is constant,[109] the probability of gaining (or losing) support for a guilty verdict varies across different nodes on the chain. The probability of gaining an additional vote for a guilty verdict depends on the current size of the guilty-verdict faction.[110] The larger the guilty-verdict faction, the higher the probability it will gain members.

This model of jury deliberation reflects essential elements of jury deliberation and avoids making unnecessary assumptions about the process.[111] This model makes no assumptions about how jurors deliberate,[112] who is elected foreperson,[113] who speaks during deliberation,[114] or why jurors change

---

109. The Gambler's Ruin problem is a classic example used to illustrate Markov chains. *See* Seongjoo Song & Jongwoo Song, *A Note on the History of the Gambler's Ruin Problem*, 20 COMMC'NS STAT. APPLICATIONS & METHODS 157, 157 (2013). In this scenario, a gambler starts with a certain amount of capital and repeatedly places bets. *See* J.L. Coolidge, *The Gambler's Ruin*, 10 ANNALS OF MATHEMATICS 181, 181–82 (1909). The game continues until the gambler either loses all their money (goes broke) or reaches a specified goal. See Coolidge supra note 107. For example, the gambler may start with $6 and repeatedly place $1 bets until he either loses all his money or reaches $12. The transition probabilities depend on the betting game and do not change depending on how much money the gambler has. The gambler might have a 50 percent probability of gaining a dollar and a 50 percent probability of losing a dollar each time he places a bet.

The Gambler's Ruin problem would describe jury deliberation without competing factions pulling persuadable jurors to their side. If p_k were constantly .5, P(G) would equal the proportion of jurors in favor of a guilty verdict.

110. *See* MacCoun & Kerr, *Asymmetric Influence*, *supra* note 100, at 30.

111. This simple Markov chain model can be summarized with one figure and estimated with a few lines of computer code. Simplicity is a virtue. Occam's Razor suggests that when there are multiple explanations of a complex phenomenon, the simplest explanation is likely to be correct. *See* Jason J. Braithwaite, *Occam's Razor: The Principle of Parsimony*, BEHAV. BRAIN SCI. CTR. 1, 1 (2007). It is more than a matter of philosophy: The model's validity is demonstrated with comparison to empirical evidence. *See* discussion *infra* Part II.C.4.

112. During deliberation, jurors typically sit around a conference table and take turns speaking. The discussion focuses on choosing a verdict or reviewing the evidence and exhibits. *See* HASTIE, PENROD & PENNINGTON, *supra* note 107, at 163–65.

113. The first juror to bring up the need to select a foreperson often hears another juror say: "So why don't you do it?" *See* Franklin J. Boster et al., *An Information-Processing Model of Jury Decision Making*, 18 COMMC'N RSCH. 524, 538 (1991); *see also* Dennis J. Devine et al., *Explaining Jury Verdicts: Is Leniency Bias for Real?*, 34 J. APPLIED SOC. PSYCH. 2069, 2080 (2004) [hereinafter *Explaining Jury Verdicts*] (finding the foreperson is most often the person who volunteers). If no one volunteers for the job, jurors typically select someone who has served on a jury before, has a high-status occupation, is the oldest person in the room, or happens to be sitting at the head of the table. *See* Phoebe C. Ellsworth, *Are Twelve Heads Better Than One?*, 52 L. & CONTEMP. PROBS. 205, 213–14 (1989); Nancy S. Marder, *Gender Dynamics and Jury Deliberations*, 96 YALE L.J. 593, 595 (1986).

The jury foreperson does not have formal powers, but the foreperson tends to speak more frequently than other jurors and may exert slightly more influence. *See* Traci Feller, *What the Literature Tells Us About the Jury Foreperson*, 22 JURY EXPERT 42, 46–47 (2010); DEVINE, *supra* note 49, at 168; Linda A. Foley & Melissa A. Pigott, *The Influence of Forepersons and Nonforepersons on Mock Jury Decisions*, 15 AM. J. FORENSIC PSYCH. 5, 12–15 (1997). *But see* Reid Hastie, David A. Schkade & John W. Payne, *A Study of Juror and Jury Judgments in Civil Cases: Deciding Liability for Punitive Damages*, 22 L. & HUM. BEHAV. 287, 296 (1998) (finding the foreperson is not more influential in civil cases) [hereinafter *A Study of Juror and Jury Judgments in Civil Cases*].

114. Juror participation is correlated with personal characteristics such as gender and socio-economic status. Jurors in the minority faction speak more frequently than those in the majority faction, as does the foreperson compared to other jurors. *See* DEVINE, *supra* note 49, at 155–56, 166–67. Some jurors speak more

their minds.[115] These factors are unnecessary complications when estimating the probability of a guilty verdict.[116] The social dynamics of deliberation are interesting, but the outcome is largely a function of the initial vote. In this model, the value of k simply increases by one, decreases by one, or remains unchanged in each round of deliberation. It is a simple model of deliberation, yet it yields precise results and aligns with the empirical record of jury deliberations.

2. *Formal Model Results*

To obtain probabilities of concluding deliberation with a guilty verdict, the model's transition probabilities are incorporated into a transition matrix.[117] Matrix algebra is then used to calculate the probabilities of reaching a guilty verdict from any initial state.[118] Additionally, the model of deliberation described earlier can be implemented as computer code and estimated using software.[119] Table 2 presents the probabilities of guilty verdicts based on the initial states of six and twelve-person juries.[120]

---

frequently than other jurors do, but a juror's willingness to speak is not correlated with voting guilty or not guilty. *Id.* at 169–70.

115. It would be problematic to elevate or discount jurors' opinions based on factors such as age, race, occupation, income, gender, or other demographic characteristics. Once jurors' initial verdict preferences are known, additional information about the jurors does not help us estimate the probability of a guilty verdict.

116. The model assumes these factors to be unnecessary complications. The validity of this assumption will be tested empirically. If the model successfully explains deliberation outcomes, additional elements become unnecessary complications.

117. A transition matrix is a square matrix that describes the probabilities of the system moving from one state to another. *See* Olivia Gandrillon et al., *Entropy as a Measure of Variability and Stemness in Single-Cell Transcriptomics*, 27 CURRENT OP. SYS. BIO. 1, 9 (2021). For example, in a six-person jury, a transition matrix shows the probabilities that there will be 0, 1, 2, 3, 4, 5, or 6 votes for a guilty verdict in the next stage of deliberation, given there are currently 0, 1, 2, 3, 4, 5, or 6 votes for a guilty verdict. There are, at most, three positive values per row because the system can do one of three things from a nonterminal state: the state of the guilty-verdict faction can increase by one, decrease by one, or stay the same.

118. *See generally* WILLIAM H. GREENE, ECONOMETRIC ANALYSIS 9–14 (4th ed. 2000) (introducing principles of matrix algebra); CARL P. SIMON & LAWRENCE BLUME, MATHEMATICS FOR ECONOMISTS 153–59 (1994) (explaining the application of matrix algebra). The transition probability matrix, P, can be partitioned into four component blocks:

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

Matrix Q contains probabilities between transient states, R contains probabilities of transitioning from transient to absorbing states, 0 is a matrix of zeros, and I is an identity matrix. The probabilities of reaching absorbing states from initial states, like those reported in Table 2, can then be solved as $B = (I - Q)^{-1} R$. The resulting B matrix contains the $P(G \mid k)$ probabilities.

119. The deliberation process is implemented by the SATE Package's "deliberate" function, which returns a "Guilty" or "Not guilty" verdict given values of k and n. *See* Edwards, *SATE*, *supra* note 12. Executed once, the deliberate function's result is random, but executing the function many times allows one to calculate precise verdict probabilities.

120. $P(G \mid k = 1, n = 12)$ is not zero; it is .0001 which rounds to .000.

Table 2. Probability of Guilty Verdict Based on Jury's Initial Poll

Initial Votes for Guilty Verdict	Probability of Guilty Verdict	
	Twelve-Person Juries	Six-Person Juries
0	0.000	0.000
1	0.000	0.004
2	0.001	0.045
3	0.009	0.206
4	0.038	0.522
5	0.119	0.839
6	0.278	1.000
7	0.501	
8	0.723	
9	0.882	
10	0.963	
11	0.993	
12	1.000	

There are several noteworthy features of the formal model results. The relationship between the jury's initial poll and the probability of a guilty verdict is a positively sloped curve.[121] Additional votes for a guilty verdict always increase the probability of a guilty verdict, but the marginal effect of an additional guilty vote varies. The formal model indicates that jury size affects verdict probabilities, with six-person juries being generally more lenient.[122] The formal model suggests that P(G) is higher when starting with a 10-2 vote in favor of a guilty verdict than with a 5-1 vote even though these are mathematically equivalent fractions. The formal model results exhibit a leniency shift consistent with the heightened burden of proof; a 7-5 vote in favor of conviction is essentially a tie (.501 probability of conviction) while a 6-6 vote favors the defendant (.278 probability of conviction).

3. Validity of the Deliberation Model

The jury deliberation model developed in this Part is useful if it generates valid and reliable results. A valid model produces unbiased estimates of the quantity it seeks to measure—in this case, $P(G \mid k)$. Previous studies of jury

121. *See* Dennis J. Devine et al., *Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups*, 7 PSYCH., PUB. POL'Y, & L. 622, 623 (2001) ("The verdict preferred by the majority of jurors on the first ballot was the jury's final verdict over 90% of the time.") [hereinafter *Jury Decision Making*].

122. It should be noted, however, that the results show that six-person juries are more likely to convict starting with a 1-5 or 2-4 initial vote in favor of a guilty verdict (compared to 2-10 or 4-8 initial votes on twelve-person juries). The differences from these starting points are mild—less than one percentage point. Assuming these starting values are not significantly more common than other values, the net effect of using six-person juries would be to lower the probability of conviction.

deliberation provide a large sample of deliberations where real or mock jurors observe a trial, form preferences about the verdict, and then deliberate in groups to reach a verdict.[123] To assess validity, we can compare the formal model's results with empirical estimates of the same quantities.[124]

Most of the data points for empirical analysis are derived from Devine et al.'s compilation of twenty-four jury deliberation studies, published in 2001.[125] I have updated that work with five additional studies: Devine et al. (2007),[126] Devine and Kelly (2015),[127] Devine et al. (2004),[128] Sandys and Dillehey (1995),[129] and Hannaford-Agor et el. (2002).[130] Data from these twenty-nine studies allow for the analysis of the relationship between juror preferences and jury verdicts based on thousands of data points.

Logistic regression is used to analyze the relationship between the probability of a binary outcome and one or more explanatory variables.[131] The outcome of interest is whether the jury returns a guilty verdict. Hung juries are excluded from the analysis.[132]

---

123. *See infra* notes 122–25; *see supra* note 120 and accompanying text.

124. For discussion of reliability, see *infra* Part II.C.4.

125. *Jury Decision Making*, *supra* note 121, at 691 tbl. 6.

126. Dennis J. Devine et al., *Deliberation Quality: A Preliminary Examination in Criminal Juries*, 4 J. EMPIRICAL LEGAL STUD. 273, 293 tbl. 4 (2007) (field study of Indiana juries).

127. Dennis J. Devine & Christopher E. Kelly, *Life or Death: An Examination of Jury Sentencing with the Capital Jury Project Database*, 21 PSYCH., PUB. POL'Y, & L. 393, 398–402 (2015). For additional analysis of the Capital Jury Project Database, see Theodore Eisenberg, Stephen P. Garvey & Martin T. Wells, *Jury Responsibility in Capital Sentencing: An Empirical Study*, 44 BUFF. L. REV. 339, 349–50 (1996); Scott E. Sundby, *War and Peace in the Jury Room: How Capital Juries Reach Unanimity*, 62 HASTINGS L.J. 103, 105 (2010).

128. *Explaining Jury Verdicts*, *supra* note 113, at 2082–83 tbl. 2, tbl. 3 (field study). I use data from this study as discussed by Kerr and MacCoun. *See* Norbert L. Kerr & Robert J. MacCoun, *Is the Leniency Asymmetry Really Dead? Misinterpreting Asymmetry Effects in Criminal Jury Deliberation*, 15 GRP. PROCESSES & INTERGROUP RELS. 585, 588 tbl. 2 (2012) [hereinafter *Leniency Asymmetry*].

129. Marla Sandys & Ronald C. Dillehay, *First-Ballot Votes, Predeliberation Dispositions, and Final Verdicts in Jury Trials*, 19 L. & HUM. BEHAV. 175, 183–85 (1995). I use data from this study as adjusted by Kerr & MacCoun, *supra* note 128, at 592 tbl. 4.

130. PAULA L. HANNAFORD-AGOR ET AL., NAT'L CTR. FOR STATE CTS. REP., ARE HUNG JURIES A PROBLEM? 33, 65–66 (2002) (evaluating 382 cases from four state court systems). The data from this study are publicly available. *See* Paula L. Hannaford-Agor et al., *Evaluation of Hung Juries in Bronx County, New York, Los Angeles County, California, Maricopa County, Arizona, and Washington, DC, 2000-2001,* NAT'L ARCHIVE OF CRIM. JUST. DATA (March 30, 2006), https://www.icpsr.umich.edu/web/NACJD/studies/3689/versions/V1 [https://perma.cc/5V6U-5FLZ].

131. In logistic regression analysis, a link function translates a linear equation into predicted probabilities of the outcome of interest. FRED C. PAMPEL, LOGISTIC REGRESSION: A PRIMER 2 (2020). The relationship between jurors' initial preferences and P(G) is not linear, but the relationship between jurors' initial preferences and the *logged odds* of a jury finding the defendant guilty may be a linear relationship. The odds of an event occurring equal the probability of the event occurring divided by the probability of the event not occurring; $odds = p / (1 - p)$ where p is the probability of the outcome of interest. JOHN FOX, APPLIED REGRESSION ANALYSIS AND GENERALIZED LINEAR MODELS 337 (2d ed. 2008). The logged odds of an outcome are a linear function of one or more explanatory variables. For comparison of logistic regression and linear probability model, see FOX, *supra*, at 335–43; PAMPEL, *supra* note 131, at 3–10.

132. Hung juries do not provide useful information for assessing the probability of a different trial outcome. If the jury cannot reach a verdict, the prosecutor may try the case again. *See* United States v. Perez, 22 U.S. 579 (1 Wheat.) 578–81 (1824) (holding that retrying a defendant after a mistrial caused by a hung jury does not violate Double Jeopardy). Although hung juries are not the focus here, the model may offer some

The key explanatory variable is the proportion of jurors who initially favor a guilty verdict.[133] Measuring the initial jury vote as the proportion of jurors who favor a guilty verdict, rather than the raw number, allows for the analysis of different jury sizes and makes efficient use of the full sample.[134] The logistic regression equation is initially estimated using only the initial jury vote to predict jury verdicts.[135]

We can further assess the validity of the Markov chain model of jury deliberation by incorporating control variables into the analysis.[136] Two variables may directly affect jury verdicts, independent of their influence on jurors' initial preferences: capital punishment and jury size.[137]

The punishment phase of a capital trial differs from the guilt phase, but it is unclear whether deliberating punishment is different than deliberating guilt; the relationship between initial preferences and verdict probabilities could be stronger, weaker, or no different than it is for guilt decisions.[138] The formal

---

insights into hung juries. To model the probability of impasse, the Markov chain model can end deliberation at a cut-off stage, rather than allow deliberation to continue indefinitely.

133. There is some debate over how to measure initial preferences when jurors abstain from the first poll or say they are undecided. Devine and his colleagues assume these jurors have reasonable doubts and prefer a not guilty verdict. *Explaining Jury Verdicts*, *supra* note 113, at 2078. Kerr and MacCoun suggest different approaches for undecided and abstaining jurors. *See Leniency Asymmetry*, *supra* note 128 at 588–91 (choosing to exclude or split these datasets instead). Kerr and MacCoun divide their votes between guilty and not guilty to measure initial support for a guilty verdict. *Leniency Asymmetry*, *supra* note 128, at 591–93. Following one suggestion of Kerr and MacCoun, I divide undecided votes and abstentions between guilty and not guilty votes where possible. The issue of uncertain or abstaining votes is limited to field studies. Researchers can require mock jurors to express their verdict preference.

134. First vote tallies in the Hannaford-Agor et al. study's cases were calculated from juror surveys. *See* HANNAFORD-AGOR ET AL., *supra* note 130, at 65. There are multiple juror surveys for each case, and sometimes jurors report differences in the tallies of their initial votes. Because multiple measures tend to reduce error, I average the jurors' reports of initial votes for the prosecution, defense, and undecideds.

135. *See infra* Table 3.

136. Variables affecting both initial preferences and deliberation may confound causal analysis and distort regression results. Some variables may influence initial preferences or the deliberation process but not both; such variables do not confound analysis of the relationship between juror preferences and jury verdicts.

137. *See* Michael J. Saks & Mollie Weighner Marti, *A Meta-Analysis of the Effects of Jury Size*, 21 LAW & HUM. BEHAV. 451, 451–52 (1997); *see also* Theodore Eisenberg, Stephen P. Garvey & Martin T. Wells, *Forecasting Life and Death: Juror Race, Religion, and Attitude Toward the Death Penalty*, 30 J. LEGAL STUD. 277, 308 (2001) [hereinafter *Forecasting Life and Death*]. The burden of proof is potentially a confounding variable, but it is held constant in criminal trials. Changing the burden of proof would likely influence both initial juror preferences as well as final jury verdicts, but the burdens of proof used in criminal jury trials do not vary from one trial to the next. *See* Norbert L. Kerr et al., *Guilt Beyond a Reasonable Doubt: Effects of Concept Definition and Assigned Decision Rule on the Judgments of Mock Jurors*, 34 J. PERSONALITY & SOC. PSYCH. 282, 293 (1976) (stating that mock jurors had a "'harshness shift' when . . . the reasonable-doubt criterion [was] stringent"). The burden of proof does, however, vary in civil trials. *See* discussion *infra* Part IV.

138. On the one hand, jurors may be reluctant to impose the ultimate punishment, potentially heightening the leniency effect in death penalty cases. On the other hand, the jury's advisory role and the judge's authority to override a death sentence recommendation in favor of life imprisonment may diminish jurors' sense of responsibility. Generally, judges cannot override jury recommendations for life imprisonment, as the Sixth Amendment requires a jury to find every fact necessary to impose a death sentence. *See* Hurst v. Florida, 577 U.S. 92, 97–99 (2016). Alabama is the only state where trial judges may override a jury's life imprisonment recommendation and impose a death sentence. *See* Patrick Mulvaney & Katherine Chamblee, *Innocence and Override*, 126 YALE L.J.F. 118, 118 (2016).

model makes no distinction between deliberating guilt and capital punishment. If studies show that deliberation is different when the issue is capital punishment, the formal model may be flawed.

Jury size has been the subject of ongoing debate and scholarly research, but the impact of using a jury with fewer than twelve members remains unclear.[139] Smaller juries deliberate differently from larger ones.[140] However, it remains unclear whether six-person juries are more likely to convict defendants.[141] Research on the impact of jury size on the probability of guilty verdicts is inconclusive.[142] For present purposes, we need not take sides in the jury-size debate, but a valid formal model of deliberation should replicate whatever effect is observed empirically.[143]

To compare the formal model of deliberation, I supply the initial conditions of the 2,303 observed jury deliberations to the Markov Chain model and have

---

It is also possible that the relationship between initial juror preferences and jury verdicts does not differ in the punishment phase of a capital trial. Prior studies suggest that death sentences are deliberated in the same manner as guilt decisions. One mock jury study reported no heightened leniency effect when jurors deliberate whether to impose a death sentence. *See* Mona Lynch & Craig Haney, *Capital Jury Deliberation: Effects on Death Sentencing, Comprehension, and Discrimination*, 33 LAW & HUM. BEHAV. 481, 491 (2009). Another study found that the jury's recommendation of a death sentence is strikingly similar to its decision to convict the defendant. *See generally Forecasting Life and Death*, *supra* note 137 at 308–09 (analyzing death penalty juries in South Carolina and finding similarity when jurors believed that "death [wa]s the only acceptable punishment for murder").

139. Jury size affects the distribution of initial verdict preferences observed on juries drawn from the jury pool, but we have already accounted for the effect of n on $P(k \mid P(g))$. *See* discussion *infra* Part II.B.1.

140. Smaller juries take less time to deliberate and are less likely to hang. *See* Saks & Marti, *supra* note 137, at 461.

141. It is particularly interesting to compare 5-1 and 10-2 votes for guilty verdicts. Although these initial splits are mathematically equivalent fractions, there may be different deliberation dynamics. There are reasons to believe a 5-1 jury will be more likely to convict than a 10-2 jury. A dissenting opinion is more likely to resist conforming to the majority opinion if joined by an ally. Two jurors, allied in dissent, may hold up better against ten jurors than one lone vote against five jurors. The Supreme Court has observed that "a person in the minority will adhere to his position more frequently when he has at least one other person supporting his argument." Ballew v. Georgia, 435 U.S. 223, 236 (1978); *see also* Robert H. Miller, *Six of One is Not a Dozen of the Other: A Reexamination of* Williams v. Florida *and the Size of State Criminal Juries*, 146 U. PA. L. REV. 621, 654 (1998) ("[A] minority viewpoint defended by an ally (the ten-to-two case) can be expected to offer significantly greater resistance to conformity pressure from the majority . . . ."). At the same time, there are also reasons to believe a 5-1 jury will be *less likely* to convict than a 10-2 jury. Five people exert less peer pressure than ten people do. People are less intimidated to speak up in small group discussions. *See generally* Nicolas Fay, Simon Garrod & Jean Carletta *Group Discussion as Interactive Dialogue or as Serial Monologue: The Influence of Group Size*, 11 PSYCH. SCI. 481 (2000) (discussing larger group size as causing less members to actively participate); Rod Bond, *Group Size and Conformity*, 8 GRP. PROCESSES & INTERGROUP RELS. 331 (2005) (describing the relationship between group size and conformity). This social dynamic may allow one to resist the influence of five more than two can resist the influence of ten.

142. *See* Joan B. Kessler, *An Empirical Study of Six-and Twelve-Member Jury Decision-Making Processes*, 6 U. MICH. J.L. REFORM 712, 715–16 (1973); J. Clark Kelso, *Final Report of the Blue Ribbon Commission on Jury System Improvement*, 47 HASTINGS L.J. 1433, 1489–90 (1996); Saks & Marti, *supra* note 137, at 453–54; Norbert L. Kerr & Robert J. MacCoun, *The Effects of Jury Size and Polling Method on the Process and Product of Jury Deliberation*, 48 J. PERSONALITY & SOC. PSYCH. 349, 354 (1985); Barbara Luppi & Francesco Parisi, *Jury Size and the Hung-Jury Paradox*, 42 J. LEGAL STUD. 399, 402–04 (2013); Adam M. Chud & Michael L. Berman, *Six-Member Juries: Does Size Really Matter*, 67 TENN. L. REV. 743, 755–57 (2000).

143. Observed deliberations in the sample are 54.8% six-person juries and 45.2% twelve-person juries.

the model predict verdicts. I then apply logistic regression analysis to the verdicts predicted by the formal deliberation model.[144] Table 3 presents the results of estimating the empirical and formal models side by side.

Table 3. Comparison of Empirical and Formal Models of Jury Deliberations

	Initial Model		Controlled Analysis	
Variable	**Empirical**	**Formal**	**Empirical**	**Formal**
Initial jury vote	11.78*** (0.50)	11.13*** (0.47)	11.63*** (0.50)	10.96*** (0.47)
Death penalty case			0.21 (0.26)	0.01 (0.24)
Six-person jury			-0.37* (0.17)	-0.73*** (0.16)
Constant	-6.87*** (0.30)	-6.91*** (0.29)	-6.56*** (0.32)	-6.35*** (0.31)
Likelihood ratio	1854.67	1753.22	1864.23	1780.33
Pseudo *R*-square	0.59	0.57	0.59	0.58
Reduction in error	73.7%	66.3%	73.7%	66.4%
Correctly classified	88.6%	86.7%	88.6%	86.7%

Notes: Sample size = 2,303; Dependent variable is the logged odds of a guilty verdict; standard errors in parentheses; * = $p < .05$, *** = $p < .001$.

Empirical analysis of deliberating juries confirms that the initial vote distribution has a profound impact on jury verdicts. The various model fit statistics reported in Table 3—the likelihood ratio, pseudo-square, proportional reduction in error, and the percentage correctly classified—are all quite impressive. The influence of the initial poll is neither new nor surprising, but empirical analysis allows us to quantify the terms of this relationship precisely, providing a benchmark for a formal deliberation model.[145] The effects of initial juror preferences derived from the formal model are statistically indistinguishable from those based on empirical analysis.

The Markov chain model does not treat death penalty deliberations differently, and empirical analysis is supportive. The coefficient for death

144. The coefficient estimates vary from one iteration to the next because the dependent variable is generated through random processes. Therefore, I generate 100 sets of simulated verdicts and average the coefficient estimates and model fit statistics.

145. Jurors' initial preferences explain jury verdicts very well. The simple model's logistic regression coefficient for initial juror preferences, 11.78, is positive and statistically significant (*P*-value < .001). The various measures of model fit are strong. This sole predictor reduces null deviance by 59%, reduces outcome prediction errors by 73.7%, and correctly classifies 88.6% of sample outcomes. For further discussion of measures of logistic regression model fit, see PHILIP H. POLLOCK III & BARRY C. EDWARDS, THE ESSENTIALS OF POLITICAL ANALYSIS 295–97 (6th ed. 2019).

penalty cases is not statistically significant, indicating that the effect of deciding a death penalty case, controlling for initial juror preferences, is indistinguishable from zero.[146]

The coefficient for six-person juries is negative and statistically significant.[147] Empirical analysis suggests that six-person juries are slightly *less likely* to convict, holding jurors' initial verdict preferences constant.[148] The Markov chain model of deliberation faithfully replicates the decrease in guilty-verdict probability with six-person juries.[149] To facilitate further comparison, Figure 4 presents the predicted probabilities of guilty verdicts based on logistic regression analysis of empirically observed and model-generated verdicts.[150]

146. This does not necessarily imply that death penalty decisions are no different than guilt decisions. The gravity of capital punishment may affect initial juror preferences, but having accounted for jurors' initial verdict preferences in these cases, there does not appear to be a direct effect of death penalty type cases on deliberation.

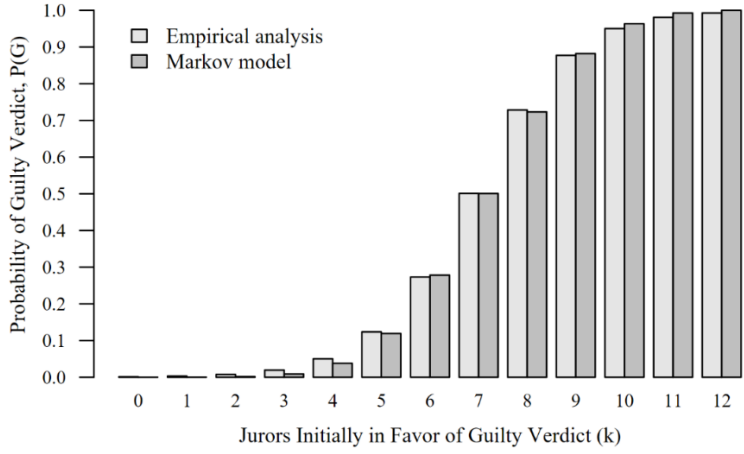
147. The results indicate that jury size has some effect on P(G) independent of, and controlling for, jurors' initial preferences. Jury size also impacts the initial distribution of juror preferences when jurors are selected from a large pool of potential jurors. *See* discussion *supra* Part II.A.

148. The difference between empirical results and the theoretical expectation that a 5-1 jury is more likely to convict than a 10-2 jury, because the lone dissenter on the six-person jury does not have "an ally," may be a byproduct of including hung juries in the analysis. According to Devine et al., of 17 twelve-person juries observed with 10-2 initial vote distributions, there were 11 guilty verdicts, 0 not guilty verdicts, and 6 hung juries. *See Jury Decision Making*, *supra* note 121, at 691 tbl. 6. Thus, the conviction rate was 68.75% if hung juries are included but 100% if hung juries are excluded. Of the 164 six-person juries observed with initial 5-1 votes, there were 136 guilty verdicts, 5 not guilty verdicts, and 23 hung juries. *Id.* Thus, the conviction rate starting from 5-1 initial votes was 82.93% including hung juries and 96.45% if hung juries are excluded. When hung juries are excluded, the conviction rate is lower starting from a 5-1 vote (96.45%) than it is starting from a 10-2 vote (100%). An ally may help a dissenting juror hold out against the majority until the judge declares an impasse, delaying but not changing the ultimate outcome.

149. The formal model suggests a stronger leniency effect using six-person juries than the empirical model does (.73 versus .37 decrease in logged odds); but the difference between coefficients for six-person juries is not statistically significant. This is an interesting result, given the policy debate over jury size, but it should not come into play very often when analyzing whether defendants receive fair trials. The cases of primary concern, such as capital crimes and felonies, are usually decided by twelve-person juries. *See* Chud & Berman, *supra* note 142, at 748 (explaining the Federal Rules of Criminal Procedure require twelve-person juries in criminal trials). No state uses juries with fewer than twelve members in capital cases. Only six states use six or eight-person juries in felony trials. *See* John Fritze, *Supreme Court Declines to Decide Whether 12-Person Juries Should be Required for State Felony Cases*, CNN (May 28, 2024, at 10:14 ET), https://www.cnn.com/2024/05/28/politics/supreme-court-juries-state-felony-cases/ [https://perma.cc/729R-PETN]. While it is possible to measure the fairness of misdemeanor trials with six jurors, it has limited practical value.

150. The logistic regression coefficients in Table 3 are reported in terms of logged odds (the natural logarithm of the odds of a guilty verdict). Logged odds are not an intuitive metric for communicating the likelihood of a guilty verdict. Fortunately, logged odds can be translated into predicted probabilities to better understand the results of logistic regression analysis. *See* POLLOCK III & EDWARDS, *supra* note 145, at 289–91; ANDREW GELMAN & JENNIFER HILL, DATA ANALYSIS USING REGRESSION AND MULTILEVEL/HIERARCHICAL MODELS 79–82 (2007).

Figure 4. Comparing Empirical Analysis and Markov Model of Jury Deliberation



The formal Markov chain model of deliberation effectively represents the relationship between the jury's initial poll and the probability of a guilty verdict. The formal model's results are nearly identical to those obtained from five decades of research on jury deliberation. For our limited purposes, the model behaves just like a real jury. It captures the essential features of jury deliberation and provides probabilities of guilty verdicts based on a jury's initial poll.

4. Reliability of the Deliberation Model

Reliability refers to the consistency and stability of a measurement, reflecting the extent to which a measurement strategy produces the same results under consistent conditions.[151] The Markov chain model of jury deliberation enhances reliability in two key respects. First, it eliminates the uncertainty inherent in empirical analyses of the relationship between the jury's initial poll and the probability of a guilty verdict. The relationship between initial juror preferences is determined mathematically rather than estimated empirically.[152] This approach yields sharper, more precise measurements, an advantage illustrated further in Part II.D.3.

Second, the results of the formal model remain stable over time. An empirical model based on past research is a moving target, subject to change,

151. Saul McLeod, *Reliability vs Validity in Research*, SIMPLYPSYCHOLOGY (Dec. 13, 2024), https://www.simplypsychology.org/reliability-or-validity.html [https://perma.cc/WT2Q-66LX].

152. Additional studies of jury deliberation would also decrease uncertainty by increasing sample size, but this is a challenging and costly endeavor. The studies analyzed *infra* Part II.C.4 are the culmination of five decades of research. Moreover, the benefits of increasing sample size are subject to diminishing returns. For instance, halving standard errors requires quadrupling the sample size. *See* ROBERT V. HOGG, JOSEPH W. MCKEAN & ALLEN T. CRAIG, INTRODUCTION TO MATHEMATICAL STATISTICS 215–16 (7th ed. 2012) (providing the definition of standard error and the central limit theorem).

revision, and ongoing debate over the inclusion or exclusion of studies and the coding of variables.[153] A formal model provides a stable and consistent metric for estimating fairness based on jurors' initial preferences.[154] The results of the formal model are durable, easy to verify, and provide a solid foundation for studying jury-level effects using juror-level data.[155]

D. Probability of a Different Result from Difference in Verdict Preferences

Given estimates of jurors' verdict preferences in the actual and hypothetical trial conditions, the process described in preceding Parts allows us to calculate P(G | actual) and P(G | hypo) based on readily obtainable, case-specific data. With these verdict probabilities, we can quantitatively assess the fairness of a defendant's trial as the difference between the probability of conviction in his actual trial and a hypothetical, error-free trial. Moreover, as detailed in this Part, we can communicate the results using graphs and account for the uncertainty of estimated values.

1. Calculating the Difference in Verdict Probabilities

To calculate guilty-verdict probabilities based on jury-pool verdict preferences, the initial poll probabilities are multiplied by corresponding guilty-verdict probabilities. The products are then summed to determine the overall probability of a guilty verdict, given the jury pool's preference. The complete formula appears below.[156]

$$P\big(G \mid P(g)\big) = \sum_{k=0}^{k=n} P(k \mid P(g)) \cdot P(G \mid k)$$

153. *See, e.g.*, *Leniency Asymmetry*, *supra* note 128, at 591–96 (recoding and reanalyzing prior studies of jury deliberation).

154. The "moving target" problem generated significant controversy in the statistical analysis of DNA test results in criminal trials. *See* ARONSON, *supra* note 11, at 26. The probability of a DNA match between a suspect and a sample relies on the estimated frequencies of genetic markers within a reference population. *Id.* at 26–29. The selection of reference database for analysis sparked controversy due to variations in databases and genetic markers, some of which were proprietary and could not be reviewed, with frequencies differing across races, ethnicities, and nationalities. *See* Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 SCI. 1745, 1745–46 (1991); Bruce S. Weir, *Population Genetics in the Forensic DNA Debate*, 89 PROC. NAT'L ACAD. SCIS. 11654, 11654 (1992); Erin Murphy & Jun H. Tong, *The Racial Composition of Forensic DNA Databases*, 108 CAL. L. REV. 1847, 1851 (2020).

155. It is also possible to generate jury-level estimates based on the deliberation process derived from empirical data. Relying on empirical analysis does not systematically alter guilty-verdict probabilities, but it does increase estimation uncertainty. *See* discussion *infra* Part II.D.3 (comparing estimation uncertainty from formal and empirical models of jury deliberation).

156. The P(G | P(g)) formula can be expressed succinctly using matrix notation. Let K represent a row matrix of probabilities of selecting 0 to k jurors who favor a guilty verdict given P(g) in the jury pool. Let M represent a column matrix of guilty-verdict probabilities given initial deliberation states. Then, P(G | P(g)) = $K \cdot M$. The formula in the text is equivalent to matrix multiplication.

To illustrate, Table 4 details the calculation of guilty-verdict probability when *King* is decided with P(g) = .760, following the *M'Naghten* rule, and P(g) = .640, following the *Durham* rule.[157] By summing the P(G | *k*) · P(*k*) values for *k* = 0, 1, 2 . . . 12, we find that the value of P(G) following *M'Naghten* is .828 and the value of P(G) following *Durham* is .617. The general formula for P(G | P(g)) allows us to calculate the probabilities of guilty verdicts in the actual and hypothetical trial conditions, the quantities compared to judge fairness, based on P(g), a value that can be readily estimated from a survey.

Table 4. Illustrative Calculation of P(G) for a Specific Value of P(g)

Initial		**P(g) = .760**		**P(g) = .640**	
Poll (*k*)	**P(G | *k*)**	**P(*k*)**	**P(G | *k*) · P(*k*)**	**P(*k*)**	**P(G | *k*) · P(*k*)**
0	.000	.000	.000	.000	.000
1	.000	.000	.000	.000	.000
2	.001	.000	.000	.001	.000
3	.009	.000	.000	.006	.000
4	.038	.002	.000	.023	.001
5	.119	.009	.001	.067	.008
6	.278	.034	.009	.138	.038
7	.501	.092	.046	.211	.105
8	.723	.183	.132	.234	.169
9	.882	.257	.227	.185	.163
10	.963	.244	.235	.099	.095
11	.993	.141	.140	.032	.032
12	1.000	.037	.037	.005	.005
P(G | P(g))			**.828**		**.617**

Notes: The P(*k* | P(g) = .760) and P(*k* | P(g) = .640) values are plotted in Figure 1. The P(G | *k*) values for twelve-person juries, obtained from the Markov Chain model of deliberation, appear in Table 2.

The probability of a different outcome is the difference between P(G | actual) and P(G | hypo).[158] Continuing with the *King* trial example, Table 4 shows that P(G | actual) = .828 and P(G | hypo) = .617. The difference in guilty-verdict probabilities represents the effect of changing the jury instruction on the outcome of *King*. Instructing the jury to follow *M'Naghten*, rather than *Durham*, increases the probability of conviction by .212.[159] Given estimates of jury-pool preferences in both the actual and hypothetical trial conditions, we

157. P(*k*) values reported as .000 are greater than zero, but round to zero with three decimal places. Table 4's P(*k*) values assume jurors are selected without peremptory strikes, but that assumption can be modified to account for peremptory strikes. *See* discussion *supra* Part II.B.2.

158. For further discussion of this measure of harm, see Edwards, *supra* note 19, at 34–37.

159. Comparing P(G) values in Table 4 (.828 and .617) yields .211 due to rounding intermediate values.

can calculate the difference in verdict probabilities to measure the fairness of criminal trials.

2. *Graphing the Probability of a Different Result*

The measure described in this Part can be visualized using graphs. Figure 5 plots the relationship between P(g) and P(G) for twelve-person juries as a solid line. This line, traced as precisely as possible, is crucial to estimating verdict probabilities from readily obtainable data.

Figure 5: Increase in the Probability of a Guilty Verdict Based on Jury Pool Preferences

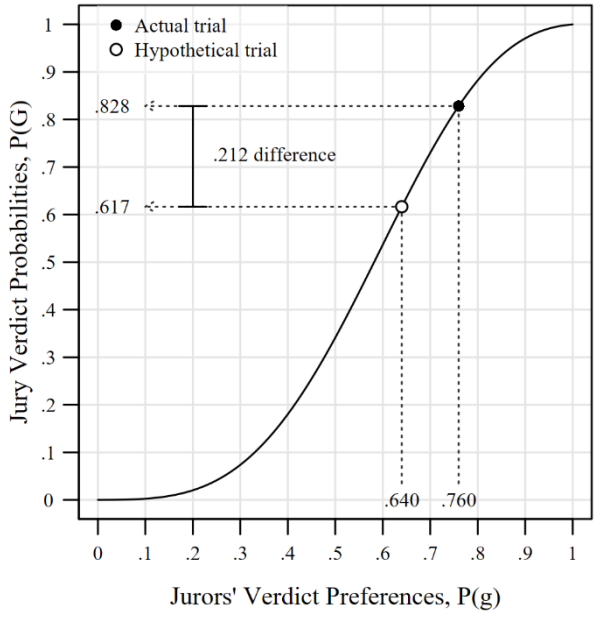


Figure 5 shows that P(g | actual) and P(g | hypo) values of .760 and .640 correspond to P(G | actual) and P(G | hypo) values of .828 and .617. Based on the guilty-verdict probabilities, the difference between the defendant's actual trial and a hypothetical, error-free trial caused a .212 increase in the probability of conviction.[160] This example highlights the importance of knowing the values of P(g) in the actual and hypothetical trial conditions, rather than simply examining the difference between them. The effect of an increase in P(g) depends on where the increase occurs.

160. An increase in P(g) will often have an outsized impact on P(G). For instance, this .120 increase in P(g) corresponds to a .212 increase in P(G). Most appeals are expected to arise in cases where jurors' initial preferences are divided rather than when nearly all jurors favor one verdict or the other. However, this is not always the case as demonstrated in the analysis of *Florida v. State infra* Part III.A and *Hopkins v. Cockrell infra* Part III.B.

How do peremptory strikes affect the relationship between P(g) and P(G), and how do they influence measures of fairness based on that relationship? Overall, the effect of strikes on harm analysis is relatively mild. In the *King* example, a D.C. trial where both the prosecution and defense are allowed ten peremptory strikes,[161] there is a .211 probability of a different outcome, assuming parties exercise strikes with realistic accuracy.[162] The estimate would differ slightly if the cases were a federal non-capital felony trial (.229),[163] a Michigan felony trial (.212),[164] or a federal death penalty trial (.206).[165] The effect of peremptory strikes on the relationship between P(g) and P(G) depends on the number of strikes available and how accurately parties can eliminate jurors likely to vote against them. Variation among jurisdictions does not present a problem in applied analysis, as the jurisdiction's rules for jury size and peremptory strikes are known constants, easily incorporated into the analysis.[166]

3. Uncertainty of Estimates

Quantifying the uncertainty of estimated values is essential. For example, if one estimates that a trial error caused a .212 increase in the probability of conviction, it is essential to convey the margin of error for that estimate. The margin of error of an estimate reflects its plausible values.[167] The greater the uncertainty in an estimated value, the wider the margin of error. [168]

There are two sources of uncertainty when estimating P(G) for a specific trial condition. There is random measurement error in the individual-level estimates of P(g) for specific trial conditions. Additionally, there may be uncertainty about the relationship between juror preferences and jury verdicts.[169] Both layers of uncertainty contribute to the overall uncertainty surrounding the value of P(G).

---

161. *See* D.C. CODE § 23-105(a) (1970).

162. Empirical research suggests that attorneys exercise peremptory strikes with ten to fifteen percent accuracy. *See supra* Part II.B.2. With ten strikes per side, P(G) is higher in both conditions, but the difference between conditions is minimal.

163. *See* FED. R. CRIM. P. 24(b)(2) (prosecution strikes six, defense strikes ten).

164. *See* MICH. COMP. LAWS § 768.12(1) (both sides allowed five strikes).

165. *See* FED. R. CRIM. P. 24(b)(1) (each side is allowed twenty peremptory strikes).

166. In other words, if the researcher is analyzing a Florida capital trial, there is no need to analyze harm as if it were a federal trial or from another state. Good estimates of P(g) should account for factors that vary among jurisdictions, such as the composition of their jury pools. P(g) estimates the verdict preferences in the relevant jury pool, so any specific conditions that justify for-cause strikes can and should be addressed when identifying the subsample of respondents who would qualify for the jury pool.

167. *See* AGRESTI & FINLAY, *supra* note 82, at 109–10; IMAI, *supra* note 82, at 326–30.

168. *See* AGRESTI & FINLAY, *supra* note 82, at 110; IMAI, *supra* note 82, at 327.

169. When the relationship between juror preferences and jury verdicts is estimated from a finite sample, the relationship is estimated with uncertainty. Regression coefficients, based on sample data, estimate the parameters of the relationship, but the true parameters (which are unknown) may be higher or lower than those estimated. The estimated regression coefficients reported in Table 3 are 11.78 and -6.87, but these estimates are accompanied by standard errors of 0.50 and 0.30, respectively.

Figure 6 illustrates the uncertainty in a P(G) estimate. Layer one, uncertainty regarding the value of P(g), is represented by the point estimate and confidence interval for jurors' verdict preferences.[170] For example, if a researcher estimates P(g) = .640 with a margin of error of .050, the plausible values of P(g) are .640 ± .050; the value of P(g) in the jury pool could plausibly be as low as .590 or as high as .690.[171] This layer of uncertainty primarily results from the limited sample size used to estimate P(g). Researchers can conduct larger surveys to reduce this margin of error, but estimation uncertainty is inherent in case-specific estimates of P(g).[172]

If one estimates P(g) for a trial condition and then uses empirical results to estimate P(G), uncertainty regarding the juror–jury relationship introduces another layer of uncertainty. The curved line in Figure 6(a) is estimated with a margin of error that reflects uncertainty regarding the relationship between initial jury polls and the probability of guilty verdicts. In addition to uncertainty about P(g) in a jury pool, there is uncertainty regarding the relationship between P(g) and P(G). The confidence interval for P(G) in Figure 7(a) has a margin of error of ±.115; its value could plausibly range from .497 to .727.

The formal model of jury deliberation reduces, but does not eliminate, estimation uncertainty; it quiets one layer of noise.[173] There remains uncertainty about the values of P(g) estimated in different trial conditions by surveying

170. Confidence intervals identify plausible uncertainty about an estimate. *See* AGRESTI & FINLAY, *supra* note 82, at 109–10. The concept is illustrated on maps that attempt to forecast a hurricane's path. The future is unknown when researchers try to forecast a hurricane's future path with different weather models. Multiple forecasts are plotted to create a "spaghetti model" that shows a plausible set of future outcomes along with a consensus forecast based on averaged forecasts. *See* Jonathan Belles, *Hurricane Spaghetti Models: Four Things You Need to Know to Track Storms Like the Pros*, THE WEATHER CHANNEL (Sept. 21, 2022), https://weather.com/science/weather-explainers/news/spaghetti-models-tropics-tropical-storm-hurricane [https://perma.cc/6NET-H2B8]. The range of plausible future paths tends to fan outward as models attempt to project further into the future. *Id.* The hurricane may defy predictions and follow an unexpected path, but the range of expected outcomes provides useful information to help people prepare and plan for the most likely outcomes. *Id.*

171. The margin of error of a sample statistic depends on the desired level of confidence, with higher confidence levels requiring wider margins of error. The 95% confidence level is typically used. *See* B. Don Franks & Schuyler W. Huck, *Why Does Everyone Use the .05 Significance Level?*, 57 RSCH. Q. FOR EXERCISE & SPORT 245, 245 (1986). 95% confidence corresponds to a 5% inferential error rate. *See* NEIL J SALKIND, STATISTICS FOR PEOPLE WHO (THINK THEY) HATE STATISTICS 156–57 (3d ed. 2008). When the estimated value of P(g) is .620 with a margin of error of ± .050, one is 95% confident that the true value of P(g) in the population is within ±.050 of .620. The error rate of a statistic estimated with a margin of error corresponding to 95% confidence is .05.

172. The margin of error of a sample proportion at the 95% confidence level is approximately equal to $1/\sqrt{n}$, where n is the survey sample size. *See* IMAI, *supra* note 82, at 332. Thus, the margin of error for an estimate of P(g) based on 100 survey respondents is approximately ±.10.

173. The gain in estimation certainty depends on the estimated value of P(g) as well as the size of the sample used to estimate P(g). When the estimated value of P(g) is .750 based on a sample size of 1,000, the formal model reduces standard errors of P(G) estimates by about 10% compared to relying on empirical analysis of jury deliberation. The percentage gain increases as the sample size used to estimate P(g) increases. When the estimated value of P(g) is .750 based on a sample size of 4,000, the formal model reduces standard errors of P(G) estimates by approximately 40%. No amount of empirical research on deliberating juries can produce equivalent gains.

individuals, but there is no additional layer of uncertainty regarding the relationship between juror preferences and jury verdicts. By creating a model of jury decision-making, rather than relying on empirical analysis, we enhance our ability to identify harmful and harmless trial errors with a reasonable degree of scientific certainty.[174]

Figure 6: Relative Uncertainty about the Value of P(G)

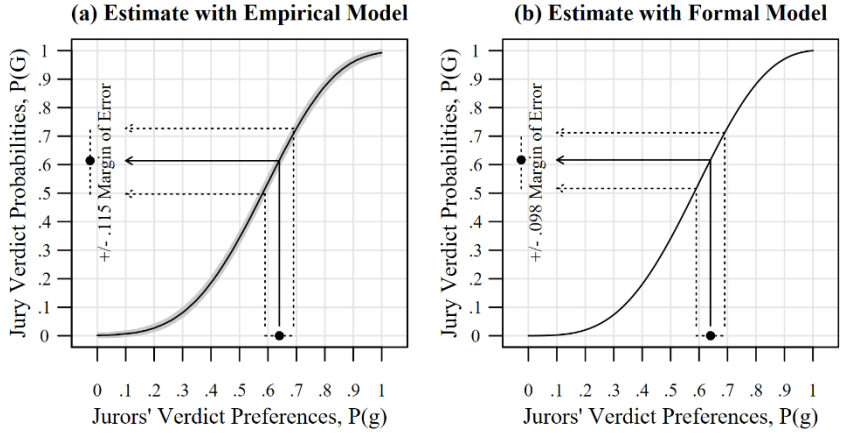


Figure 6(b) shows that while there is still uncertainty about the value of P(g), there is no additional margin of error surrounding the curved line. P(G) has a margin of error of ±.098 due to the range of plausible P(g) values. To further narrow the range of plausible P(G) values, the researcher must estimate P(g) with a larger sample.[175]

Accounting for uncertainty is necessary to avoid overconfidence, but random errors obscure authentic signals and lead to inconclusive results.[176] Uncertainty makes it more difficult for both parties to meet their burdens of proof with respect to the harmfulness of trial errors.[177] Greater certainty regarding the value of P(G) increases the power of the analysis, enabling both

174. Table 3 shows standard errors for logistic regression analysis of the formal model's predicted verdicts. Estimation uncertainty in this instance is the result of limiting the sample size to allow for side-by-side comparison with empirical data. *See infra* Table 3. When we calculate Table 4's probabilities, we are not restricted to 2,303 observations.

175. Certainty about the value of P(G) does not mean that jury verdicts are predictable. Jury verdicts remain probabilistic. We just know precisely what those probabilities are. If the jury's initial poll is 7-5 in favor of conviction, the jury's verdict is nearly as unpredictable as a coin flip. According to Table 2, $P(G \mid k = 7, n = 12) = .501$. However, one can pinpoint the probability of a guilty verdict with a high degree of certainty.

176. *See* AGRESTI, *supra* note 82, at 114 ("In forming a confidence interval, we compromise between the desired confidence that the inference is correct and the desired precision of estimation. As one gets better, the other gets worse.").

177. The burden of proof in harmless error analysis "depends on the procedural context." *See* Edwards, *Scientific Framework*, *supra* note 13, at 23. On direct appeal, the prosecution must prove that an error was harmless; in post-conviction proceedings, the petitioner must prove that an error was harmful. *See id.* at 23–26.

parties to test hypotheses about the harmfulness of trial errors more efficiently without favoring either side.[178] Figure 6 shows that the estimated value of P(G) rounds to .62 using either model of deliberation. The gain may appear modest, but it proves helpful in close cases where precise analysis offers the most benefit.[179]

The uncertainty of an estimate of harm can be visualized through graphs using confidence intervals. To illustrate, consider Simon's study of the *King* trial, where P(g | actual) = .760 and P(g | hypo) = .640, estimated with samples of 240 and 312 jurors, respectively. As discussed earlier, the difference between these trial conditions corresponds to a .212 increase in the probability of a guilty verdict.[180] Figure 7 illustrates the difference between the actual and hypothetical trial conditions, estimated with uncertainty. The ellipses in Figure 7 represent confidence intervals, identifying the range of plausible values given the uncertainty of estimating P(g) from finite samples.[181]

Uncertainty regarding P(G) in the actual and hypothetical trial conditions carries over to uncertainty about the difference between the two estimated values. The estimated probability of a different result (.212) is plotted in the right margin of Figure 7, along with its plausible range of values, based on the estimate's margin of error.[182]

178. It is possible to use empirical analysis, rather than the formal model, to derive estimates of P(G) from P(g). As discussed in the text, this approach generates duller estimates of P(G) and is more likely to yield inconclusive results. *See supra* Part II.C.4. Statistical power is the probability of correctly rejecting a false null hypothesis as opposed to failing to reject a false null hypothesis due to small sample size or standard measurement errors. *See* GELMAN & HILL, *supra* note 150, at 439–43 (discussing relationship between standard measurement error and statistical power).

179. *See* discussion *infra* Part III.D (referencing Simon's analysis of *U.S. v. King*). Greater precision allows us to demonstrate that altering jury instructions would cause intolerable harm in the *King* case.

180. *See* discussion *supra* Parts II.D.1–2.

181. The confidence intervals for P(g) and P(G) define the axes of an ellipse centered on the point estimates of P(g) and P(G). The ellipse is rotated so that its primary axis is parallel to the line relating P(g) to P(G).

182. The margin of error of the difference between estimates reflects uncertainty regarding the values being compared. If the values being compared have margins of error denoted as a and b, the margin of error of the estimated difference is $\sqrt{a^2 + b^2}$. *See* AGRESTI *supra* note 82, at 183, 185–86, 191; POLLOCK & EDWARDS, *supra* note 145, at 208–15. Applied to the current example, the margin of error of the difference between P(G | actual) and P(G | hypo) reflects uncertainty regarding the values being compared. Figure 6(b) shows the procedure for finding the margins of error of P(G | P(g = .70)) and P(G | P(g | .62)). The values being compared have margins of error of .060 and .071 when P(g) values are estimated with 500-person samples. Thus, the margin of error of the estimated effect of trial error is equal to $\sqrt{.060^2 + .071^2} = .093$. The margins of error are represented by confidence intervals in Figure 7. *See infra* Figure 7.

Figure 7: Probabilities of Different Outcomes with Margins of Error

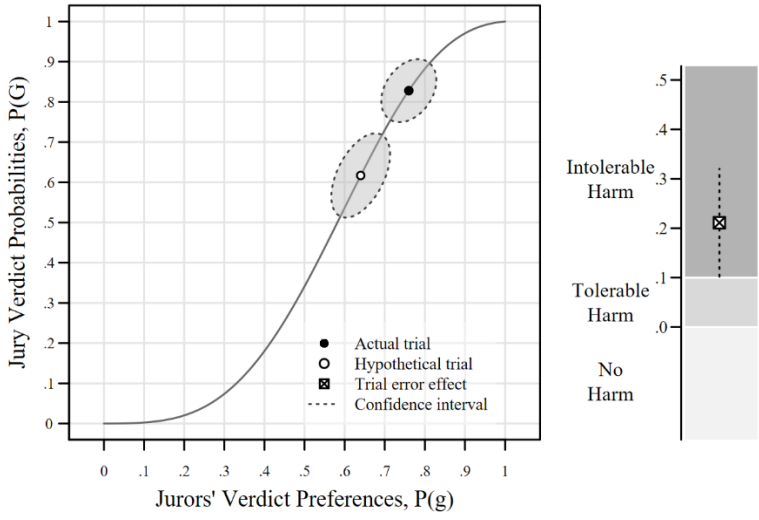


Figure 7 communicates the difference between the defendant's actual trial and a hypothetical, error-free trial. The difference resulted in a .212 increase in the probability of conviction, with ±.110 margin of error. The analysis quantifies the probability of a different outcome with reasonable certainty.[183] Is this "a probability sufficient to undermine confidence in the outcome?"[184] The Supreme Court has not yet quantified the threshold between tolerable and intolerable harms; nevertheless, .100 is a strong candidate because it is thought to be sufficient to create reasonable doubt in the minds of jurors and undermine the confidence necessary to find a defendant guilty.[185] Although .100 may be a probability sufficient to undermine confidence, it is not particularly important that .100 is the threshold, rather than another number, only that there is some identifiable standard differentiating fair from unfair.[186]

183. The plausible range of the estimated effect is between .102 and .322. *See supra* Figure 7.

184. *See* Strickland v. Washington, 466 U.S. 668, 694 (1984).

185. *See A Scientific Framework*, *supra* note 13, at 22–23, 43–44. Research suggests that reasonable doubt corresponds to approximately .10 probability in the minds of jurors. *See* Mandeep K. Dhami, Samantha Lundrigan & Katrin Mueller-Johnson, *Instructions on Reasonable Doubt: Defining the Standard of Proof and the Juror's Task*, 21 PSYCH., PUB. POL'Y, & L. 169, 169, 175 (2015); Francis C. Dane, *In Search of Reasonable Doubt*, 9 LAW & HUM. BEHAV. 141, 142–44 (1985). *But see* Rita James Simon, *"Beyond a Reasonable Doubt"— an Experimental Attempt at Quantification*, 6 J. APPLIED BEHAV. SCI. 203, 203 (1970). Reasonable doubt is a possibility sufficient to undermine confidence about a guilty verdict; thus, .100 can be viewed as "a probability sufficient to undermine confidence in the outcome" of a criminal trial. *See Strickland*, 466 U.S. at 694.

186. Identifying the probability of a different outcome is analogous to pinpointing the location of a pitched baseball. Precise measurement does not tell us whether the quantity is "too high" based on the rules of the game. I use a .100 probability of a different outcome as a thoughtful default setting to demonstrate applied analysis. There simply needs to be a standard. If a "reasonable probability" of a different outcome cannot be defined in a meaningful way that permits judgment, appellate judges are left in the unfortunate position of umpires asked to call balls and strikes with no defined strike zone.

The method discussed in this Part allows us to estimate the change in probability of conviction based on a study of juror preferences. Given the results reported by Simon, instructing jurors to apply the *M'Naghten* rule, rather than the *Durham* rule, in *U.S. v. King* would increase the probability of conviction by .212 ± .110. If this trial error is considered on direct appeal, the results do not prove that the error was harmless.[187] If the trial error is reviewed in a post-conviction proceeding, where the petitioner bears the burden of proving that the defects in his trial were harmful, the results show, with reasonable certainty, that the trial error was harmful. The error's lowest plausible effect on the probability of conviction, .102, exceeds the suggested threshold for tolerable harm.[188]

The preceding discussion explained how to calculate the probability of a guilty verdict based on the proportion of jurors who favor a guilty verdict in a modest-sized sample.[189] With modern survey technology, this analysis can be completed in a matter of weeks, if not days.[190] Identifying the relationship between juror-level data and jury-level outcomes allows for the effective estimation of the harmfulness of trial errors, potentially solving a problem that has perplexed appellate courts for centuries.[191] It is possible to estimate an important quantity—the fairness of trials—which previously appeared immeasurable. In the next Part, I apply the preceding analysis to real criminal trials to demonstrate that trial fairness can be measured using readily available data.

187. As noted earlier, Simon's *King* study addresses the effect of jury instructions, not trial errors per se. *See* SIMON, *supra* notes 56–59 and accompanying text. Because there was no mistaken instruction in King's actual trial, there is no appellate opinion assessing the harmfulness of a mistaken instruction in his case. *See* SIMON, *supra* note 56 at 50.

188. Because the harmfulness of a trial error is estimated with uncertainty, the estimated increase in probability of conviction must be substantially greater than .10 to conclude the harm was intolerable with reasonable certainty. *See A Scientific Framework*, *supra* note 13 at 44–45 ("In some cases, however, the analysis will be inconclusive because the CI of the harm estimate crosses over T = .100."). For the *King* case, the effect of a mistaken instruction is estimated to be a .212 increase in the probability of conviction, but the lower bound of its confidence interval (.102) barely clears the .100 threshold. *See* SIMON, *supra* note 56, at 64.

189. Another advantage of a formal deliberation model over empirical analysis is the ability to model different jury practices. Like using a crash test dummy to study vehicle safety, we may understand the effect of procedural changes long before they become apparent through painful lived experiences. *See, e.g.*, Barry C. Edwards, *Evaluating Death Sentencing Procedures: Fairness, Disparities, and Constitutional Limits*, B.U. PUB. INT. L.J. (forthcoming 2026) (estimating the effect of procedural variations, such as non-unanimous verdict rules, varying responses to hung juries, and judicial override procedures, on death sentence probabilities).

190. *See* Jeremy C. Wyatt, *When to Use Web-based Surveys*, 7 J. AM. MED. INFORMATICS ASS'N 426, 427 (2000) (highlighting the relationship between the mechanism of "data capture" and the speed with which it can be collected and analyzed).

191. *See* ROGER J. TRAYNOR, THE RIDDLE OF HARMLESS ERROR 4 (1970) (examining "that most pervasive and elusive of all problems in an appellate court, the riddle of harmless error").

III. MEASURING THE FAIRNESS OF REAL CRIMINAL TRIALS

Studies based on real criminal trials serve as test cases for the method of measuring fairness discussed in this Article. This method allows us to measure fairness, classify the trials as fair or unfair, and compare these assessments with those made independently by appellate courts. The method should produce analyses that are generally consistent with appellate court opinions across different types of cases.

A. Ineffective Assistance of Defense Counsel

This Part discusses two studies conducted by the author that applied the methods discussed in this Article to a pair of post-conviction cases in which petitioners claimed they were harmed by the omission of mitigation evidence. The cases share many similarities, yet the U.S. Supreme Court held that the omission of mitigation evidence was harmful in one case and harmless in the other.

George Porter, a Korean War veteran, murdered his ex-wife and her boyfriend.[192] In 1987, he pleaded guilty to two counts of first-degree murder.[193] Porter's defense counsel conducted a cursory investigation into his client's background before sentencing,[194] supposedly due to Porter's fatalism and lack of cooperation.[195] Porter was subsequently sentenced to death.[196]

After seeking relief from state courts,[197] Porter petitioned for a federal writ of habeas corpus, claiming he had been deprived of effective assistance of counsel during the sentencing phase of his trial.[198] Porter's attorney failed to present evidence of his client's abusive childhood and service-related injuries sustained during the Korean War.[199] The central issue was whether ineffective assistance of counsel created a reasonable probability of a different outcome.[200]

---

192. *See* Porter v. McCollum, 558 U.S. 30, 30–32 (2009).

193. Porter v. State, 564 So. 2d 1060, 1061–62 (Fla. 1990), *cert. denied*, 498 U.S. 1110 (1991).

194. *See Porter*, 558 U.S. at 39–40 (describing counsel's inaction).

195. *Id.* at 40. Porter represented himself at trial with standby counsel and pleaded guilty. *See Porter*, 564 So. 2d at 1061–62.

196. *Porter v. State*, 564 So. 2d at 1061–62.

197. Porter v. State, 788 So. 2d 917 (Fla. 2001); Porter v. Crosby, 840 So. 2d 981 (Fla. 2003).

198. *See Porter v. McCollum*, 558 U.S. at 38.

199. *Id.* at 33–35.

200. *See id.* at 38–39. While on Florida's death row, Porter moved for post-conviction relief in state court in 1995. *See* Porter v. State, 788 So. 2d 917, 919–20 (Fla. 2001) *cert. denied*, 534 U.S. 1004 (2001). The state trial court summarily denied all of Porter's claims except for his ineffective assistance of counsel claim, which it denied after an evidentiary hearing; Florida's Supreme Court also denied Porter post-conviction relief. *See id.* at 920, 925, 928. Porter then filed a petition for a writ of habeas corpus in the Florida Supreme Court, which again denied Porter's claims. *See Porter v. Crosby*, 840 So. 2d at 983, 986.

After exhausting state court remedies, Porter filed for a writ of habeas corpus in the Middle District of Florida, which granted him relief based on his ineffective assistance of counsel claim. *See* Porter v. Crosby, No. 603-cv-1465-Orl-31, 2007 WL 1747316, at *2–3, *31–32 (M.D. Fla. June 18, 2007). The Eleventh Circuit

The U.S. Supreme Court determined that the omission of mitigation evidence was harmful.[201]

Similar to Porter, David Leeroy Washington pleaded guilty to multiple murders in Florida, was sentenced to death, and petitioned for writ of habeas corpus, arguing that his trial counsel failed to present mitigation evidence during sentencing.[202] Compared to *Porter*, the mitigation evidence in *Washington* was relatively weak, and the aggravating factors were stronger.[203] Additionally, the Supreme Court noted that calling character witnesses could have backfired on Washington.[204] If he introduced evidence of his good character, he would have opened the door for the prosecution to present evidence of his bad character.[205] The U.S. Supreme Court ultimately decided that the omission of Washington's mitigation evidence was not harmful.[206]

In both the *Porter* and *Washington* studies, respondents are randomly assigned to either an actual trial condition with limited mitigation evidence or a hypothetical trial condition with the mitigation evidence that was omitted by

---

reversed, reasoning that the Florida Supreme Court was owed greater deference. *See* Porter v. Att'y Gen., 552 F.3d 1260, 1274–75 (11th Cir. 2008), *rev'd per curiam*, Porter v. McCollum, 558 U.S. 30, 38 (2009). The U.S. Supreme Court granted certiorari and reversed the Eleventh Circuit in 2009, granting Porter relief from his death sentence, twenty-two years after it was imposed. *See Porter v. McCollum*, 558 U.S. at 38. According to the Florida Department of Corrections, Porter died in prison in 2016. *See Inmate Release Information Detail*, FLA. DEP'T OF CORR. (Feb. 20, 2016), http://www.dc.state.fl.us/offenderSearch/detail.aspx?Page=Detail&DCNumber=110825&TypeSearch=IR [https://perma.cc/3GBE-AXEM].

201. *See Porter v. McCollum*, 558 U.S. at 44.

202. Washington pleaded guilty to murder and was sentenced to death in 1976. *See* Washington v. State, 362 So. 2d 658, 660, 662 (Fla. 1978), *cert. denied*, 441 U.S. 937 (1979) (describing the lower court proceedings). His defense attorney did not give people who would have offered mitigating evidence the opportunity to testify. *See* Strickland v. Washington, 466 U.S. 668, 672–73 (1984). The Florida Supreme Court upheld the conviction and sentence. *See* Washington v. State, 362 So. 2d at 667. In 1980, Washington filed an unsuccessful motion for post-conviction relief in the state trial court and was denied clemency by the Governor of Florida. *See* Washington v. State, 397 So. 2d 285, 286 (Fla. 1981). The Florida Supreme Court upheld the denial of relief. *See id.* at 287.

After exhausting his avenues for relief in state courts, Washington turned to the federal courts, filing an unsuccessful petition for a writ of habeas corpus in the Southern District of Florida. *See* Washington v. Strickland, 673 F.2d 879, 882, 884–85 (5th Cir. 1982) *rev'd en banc*, 693 F.2d 1243 (5th Cir. 1982) (describing the procedural history of the case). In 1982, the U.S. Court of Appeals for the Fifth Circuit, rehearing the case en banc, vacated and remanded the district court's decision. *See Washington v. Strickland*, 693 F.2d at 1246 (en banc), *rev'd*, 466 U.S. 668 (1984). The U.S. Supreme Court granted certiorari and reversed the Fifth Circuit in 1984. *See Strickland v. Washington*, 466 U.S. at 684, 701, *reh'g denied*, 467 U.S. 1267 (1984). The Supreme Court rendered its decision only eight years after Washington was sentenced to death. Washington filed a second unsuccessful petition for writ of habeas corpus in 1984. *See* Washington v. Wainwright, 587 F. Supp. 525, 526 (S.D. Fla.), *aff'd*, 737 F.2d 922 (11th Cir. 1984). Washington was executed in 1984. *See* Jesus Rangel, *Confessed Murderer of 3 Executed in Florida*, N.Y. TIMES, July 14, 1984, at 24.

203. *See Strickland v. Washington*, 466 U.S. at 699–700 (noting the "utterly overwhelming" aggravating factors).

204. *Id.* at 700.

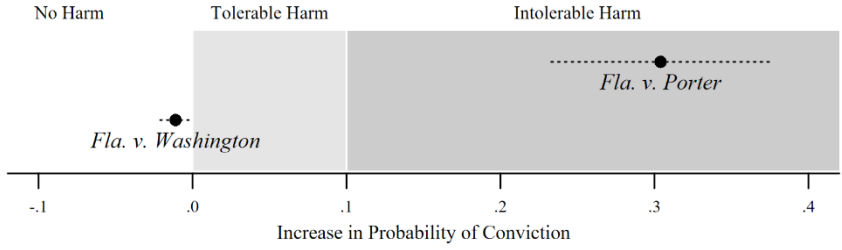
205. *Id.* The prosecution's rebuttal evidence of bad character, as detailed in the appellate record, is incorporated into the hypothetical trial vignette in the survey conducted. *See* Edwards, *Survey Results*, *supra* note 63, at *Washington survey results.tab*.

206. *Strickland v. Washington*, 466 U.S. at 700.

ineffective counsel. Table 1 reports respondents' sentence preferences in the actual and hypothetical versions of the *Porter* and *Washington* trials.

Having estimated sentencing preferences, I apply the methods described in Parts II.B, II.C, and II.D to estimate how the mitigation evidence that ineffective counsel omitted affected the probability that Florida juries would sentence Porter and Washington to death.[207] The estimates are displayed in Figure 8.[208]

Figure 8. Effect of Ineffective Defense Counsel on Probability of Death Sentence



Based on this analysis, the mitigation evidence omitted from Porter's trial increased the probability that he would be sentenced to death by .304 ± .071. The lowest plausible amount of harm is greater than the suggested threshold of tolerable harm. In other words, even the lowest plausible increase in the probability of a death sentence exceeds the threshold of tolerable harm. This analysis aligns with the Supreme Court's finding that the error in Porter's trial was harmful.

Comments from survey respondents further support the quantitative analysis of the effect of ineffective defense counsel in Porter's case:[209] "I changed my mind solely on his involvement in the war;" "PTSD from serving in Vietnam was my reason for not suggesting the death penalty;" "[T]he information . . . humanizes the defend[a]nt;" "If I never heard about those experiences, I would assume he was merely a cold-blooded killer."[210] Respondents indicated that the mitigation evidence of childhood trauma and service-related injuries affected their perceptions of Porter and the punishment he deserved.[211] Even respondents who continued to vote for the death penalty were moved: "The child abuse and brain damage makes me think the crimes was less 'his fault;'" "[I] haven[']t changed my mind but [I] do feel a little bit

207. Florida employs twelve-person juries in capital trials and allows each side 10 peremptory challenges. *See* FLA. R. CRIM. P. 3.270, 3.350(a)(1).

208. *See* Edwards, *Survey Results*, *supra* note 63.

209. One hundred twenty-one respondents volunteered comments in the survey. *See* Edwards, *Survey Results*, *supra* note 63.

210. *See id.*

211. *See id.*

sorry for him."[212] In Porter's case, the admission of mitigation evidence changes some minds and significantly affects the probability of a death sentence.[213]

In contrast to Porter's case, the omission of mitigation evidence by Washington's attorney did not cause significant harm. The estimated effect of the omission was a .011 *decrease* in the probability of a death sentence.[214] The analysis suggests that omitting evidence of Washington's "good character" may have lowered the probability that he would be sentenced to death.[215]

The harmlessness of the omitted mitigation evidence in Washington's case is reflected in the assessments offered by survey respondents. Respondents still believe that aggravating factors outweigh the mitigation evidence: "It doesn't matter. He killed multiple innocent people;" "The crimes were heinous and terrible. The victims suffered."[216] To some respondents, the mitigation evidence makes Washington appear even worse and more deserving of the death sentence: "This makes my mind more set on the decision that the death penalty is the right decision for this man. His motives were selfish;" "I feel even more strongly now that he needs to be sentenced to death as soon as possible."[217]

Based on the confidence interval for the effect of trial error on the defendant, one can confidently conclude that any harm caused by the trial error was tolerable. From this analysis, Washington would be unable to carry his burden of proving that the trial error was harmful.[218] This result is consistent with the U.S. Supreme Court's analysis of the case.[219]

B. Trials with Coerced Confessions

This Part applies the Article's method of measuring fairness to two cases where defendants appealed murder convictions due to the use of their coerced confessions against them at trial. In both cases, the central question was whether the prosecution could prove that the improper admission of confession evidence was harmless. The U.S. Supreme Court sided with one defendant, while the U.S. Court of Appeals for the Fifth Circuit ruled against the other.

---

212. *See id.*

213. *See id.*

214. The results suggest Washington's mitigation evidence would have opened the door to damaging rebuttal evidence. They show, with reasonable certainty, that the omission lowered the probability of a death sentence—an interesting result, but not the relevant inquiry for Washington's petition.

215. In Washington's case, the trial error is estimated to have *decreased* P(g) from .911 to .879. *See supra* Table 1. The resulting effect on P(G) was a decrease from .989 (trial with error) to .978 (trial without error).

216. *See* Edwards, *Survey Results*, *supra* note 63.

217. *Id.*

218. Based on the survey responses, a Florida jury is likely to recommend a death sentence for Washington with or without the mitigation evidence. *See id.*

219. *See* Strickland v. Washington, 466 U.S. 668, 699–701 (1984) (holding that the trial error was not harmful since the mitigation evidence was generally not helpful to the defendant's case).

An Arizona jury found Oreste Fulminante guilty of murdering his stepdaughter in 1985.[220] At trial, the prosecutor called two witnesses to testify that the defendant had confessed to killing his stepdaughter.[221] The first witness, Anthony, testified that Fulminante confessed to him while they were both incarcerated.[222] The second witness, Donna, testified that Fulminante confessed to her during a car ride.[223] The case against Fulminante, apart from the confessions, was circumstantial and relatively weak.[224] The confession to Anthony was later deemed inadmissible, but Donna's testimony remained admissible.[225] The U.S. Supreme Court ultimately decided that the state could not prove the error was harmless.[226]

Bobby Ray Hopkins was convicted of murdering two women in Texas.[227] The prosecution presented significant physical evidence linking Hopkins to the

220. Oreste Fulminante's eleven-year-old stepdaughter, Jeneane, was found dead in the desert around Mesa, Arizona in 1982. *See* State v. Fulminante, 778 P.2d 602, 605 (Ariz. 1988). Fulminante was convicted of murdering her in 1985, but on direct appeal, the Arizona Supreme Court, sitting in banc, ruled that the admission of his involuntary confession was reversible error and ordered a new trial. *Id.* at 606, 627. In 1990, the U.S. Supreme Court granted certiorari and, in 1991, affirmed, on different grounds, the Arizona Supreme Court's decision to order a new trial. *See* Arizona v. Fulminante, 494 U.S. 1055 (1990) (granting certiorari); Arizona v. Fulminante, 499 U.S. 279, 282 (1991).

Fulminante was retried for Jeneane's murder without evidence of his coerced confession and was again found guilty. *See* State v. Fulminante, 975 P.2d 75, 81 (Ariz. 1999). In the retrial, the prosecution introduced inadmissible hearsay statements from the victim. *Id.* In 1999, the Arizona Supreme Court, sitting in banc, found Fulminante's second trial defective and ordered yet another trial. *Id.* at 94. Rather than face a third murder trial, Fulminante pled guilty to kidnapping and second-degree murder in 1999. *See Criminal Court Case Information - Case History*, JUD. BRANCH OF ARIZ., MARICOPA CNTY., http://www.superiorcourt.maricopa.gov/docket/CriminalCourtCases/caseInfo.asp?caseNumber=CR0000-142821 [https://perma.cc/Z6H8-8NG9] (last visited Sept. 30, 2024). He was sentenced to thirteen years for kidnapping and twenty years for second degree murder. *See* E-mail from Debbie MacKenzie, Maricopa County Attorney's Office, Custodian of Records (Aug. 8, 2019) (on file with author).

221. State v. Fulminante, 778 P.2d 602, 606 (Ariz. 1988).

222. *Id.*

223. *Id.*

224. *See Arizona v. Fulminante*, 499 U.S. at 296–300.

225. *Id.* at 283, 284 & n.1.

226. The U.S. Supreme Court's decision in *Arizona v. Fulminante* is fractured into three parts. *Arizona v. Fulminante*, 499 U.S. at 282, 302–03. The Court first decides by a 5-4 vote, with Justice Scalia voting with four liberal Justices, that Fulminante's confession to Anthony in prison was coerced. *Id.* at 282 n.†, 287. Next, the Court decides, with Justice Kennedy joining four conservative Justices, that the erroneous admission of the confession at trial should be subject to harmless error analysis, not automatic reversal. *Id.* at 312 (Rehnquist, C.J., dissenting in part); *see id.* at 313 (Kennedy, J., concurring in part). Finally, the Court concludes, by another 5-4 margin, that Arizona cannot prove the confession evidence was harmless. *Id.* at 296 (majority opinion). Four of the five Justices who supported this conclusion opined that coerced confessions are inherently harmful, should result in automatic reversal, and should not be subject to harmless error analysis. *Id.* at 295 (White, J., dissenting in part). Only Justice Kennedy held both that harmless error analysis should apply and that it favored Fulminante. *Id.* at 313 (Kennedy, J., concurring in part). Kennedy did not conclude that the error was harmful; rather, he found that Arizona failed to prove the error was harmless. *Id.* at 296. Four out of five Justices, who endorsed applying harmless error analysis in coerced confession cases, found the error in Fulminante's trial to be harmless, reasoning that Donna testified to the same basic facts as her husband. *Id.* at 312 (Rehnquist, C.J., dissenting in part).

227. *See* Hopkins v. Cockrell, 325 F.3d 579, 580–82 (5th Cir. 2003) *cert. denied sub nom.*, Hopkins v. Dretke, 540 U.S. 968 (2003) (mem.).

murders, but also introduced a coerced confession at his trial.[228] The police had tricked Hopkins into confessing to the murders while he was held in jail.[229] The case against Hopkins, apart from confessions, was strong and could be made using physical evidence.[230] The Fifth Circuit determined that the admission of the coerced confession at trial was a harmless error.[231]

In the *Fulminante* and *Hopkins* studies, respondents were randomly assigned to either an actual trial condition with a coerced confession or a hypothetical trial condition without a coerced confession. After reading trial summaries, respondents were asked whether they believe the defendant should be found guilty or not guilty.

Juror verdict preferences in the *Fulminante* and *Hopkins* cases (reported in Table 1) were used to estimate the effect of coerced confession on the probability of conviction in these cases using the method outlined in Part II, including state-specific rules for jury size and the use of peremptory strikes.[232] The analysis indicates that the prison informant who improperly testified that Fulminante confessed to murdering his stepdaughter increased the probability of conviction by .086 ± .073.[233] Results in the *Hopkins* case indicate that the

228. *Id.* at 584–85.

229. Hopkins committed the murders in 1993 and was convicted in Jackson County, Texas in 1994. *Id.* at 580–82. The Texas Court of Criminal Appeals upheld his conviction and death sentence in 1997. *See* Hopkins v. Cockrell, No. 98–CV–2355–P, 2001 WL 1167312, at *1 (N.D. Tex. Sep. 28, 2001). In 1998, Hopkins filed a state habeas corpus petition, which was denied. *Id.* The following year, he petitioned the Northern District of Texas for habeas relief, but the district court, adopting a magistrate's recommendation, denied the petition in 2001. *See id.* at *1, *24. The Fifth Circuit disagreed with the District Court on the voluntariness of Hopkins' confession but affirmed the habeas denial, finding the error harmless. *See Hopkins*, 325 F.3d at 586 (5th. Cir. 2003). After his unsuccessful federal habeas proceedings, Hopkins filed another state habeas petition, challenging the state's method of execution, which was also unsuccessful. *See Ex parte* Hopkins, 160 S.W.3d 9, 9 (Tex. Crim. App. 2004) (mem.), *cert. denied*, 540 U.S. 1173 (2004). Hopkins was executed by lethal injection in 2004. *See* Michael Graczyk, *Convicted Killer in Two Knife Slayings Executed*, MIDLAND REP.-TEL. (Feb. 11, 2004), https://www.mrt.com/news/article/Convicted-killer-in-two-knife-slayings-executed-7748006.php. [https://perma.cc/9946-2W79].

230. *See Hopkins*, 325 F.3d at 585 (5th Cir. 2003).

231. *See id.* at 586.

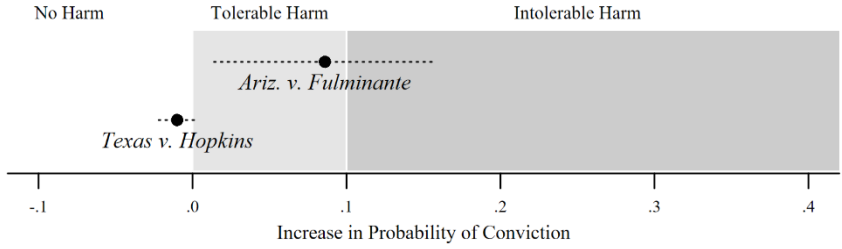
232. Arizona requires twelve jurors in capital cases. *See* ARIZ. REV. STAT. ANN. § 21-102 (2022). At the time of Fulminante's appeal, Arizona allowed each side to exercise ten peremptory challenges in trials for capital crimes. *See* ARIZ. R. CRIM. P. 18.4(c)(1)(A) (repealed 2021). As of January 1, 2022, Arizona does not permit peremptory strikes in criminal trials. *See* Order Amending Rules 18.4 and 18.5 of the Rules of Criminal Procedure, and Rule 47(e) of the Rules of Civil Procedure, No. R-21-0020 (Ariz. Aug. 30, 2021); Cheryl Corley, *Arizona's Supreme Court Eliminates Peremptory Challenges*, NPR (Sept. 6, 2021, at 05:04 ET), https://www.npr.org/2021/09/06/1034556234/arizonas-supreme-court-eliminates-peremptory-challenges [https://perma.cc/9NFC-M335]. The effect of the coerced confession is analyzed under Arizona's old rule for peremptory strikes to facilitate comparison to the Supreme Court's analysis of the case.

Texas also requires twelve jurors in felony trials. *See* TEX. CODE CRIM. P. ANN. art. 33.01 (West 2025). When the prosecution seeks the death penalty, as it did in *Hopkins*, Texas allows both the state and the defendant to exercise fifteen peremptory challenges. *See id.* at art. 35.15.

233. *See* Edwards, *Survey Results*, *supra* note 63, at *Fulminante Survey Data.tab*. The estimated values of P(G | actual) and P(G | hypo) are .282 and .196, respectively. *Id.* The probability of a guilty verdict in either condition is relatively low. In both conditions, respondents were reluctant to convict without physical evidence linking the defendant to the crime, with or without the confession testimony from a jailhouse informant. *See id.*

improper admission of the coerced confession decreased the probability of conviction by .010 ± .011.[234] Figure 9 shows the results in both cases.

Figure 9. Effect of Coerced Confessions on Probability of Conviction



The improper use of Hopkins's involuntary confession did not render his trial unfair. The addition of a questionable confession to otherwise strong forensic evidence may have actually weakened the state's case.[235] The upper bound of the confidence interval is less than the threshold of tolerable harm. In this appeal, the state bears the burden of proving the error was harmless, and based on these results, one can confidently conclude that the error was indeed harmless. This result aligns with the Fifth Circuit's harmless error analysis.[236]

Analysis of Fulminante's trial does not prove, to a reasonable degree of certainty, that the erroneous admission of an involuntary confession was harmless. The upper bound of the confidence interval, .159, exceeds the suggested threshold of tolerable harm.[237] This result aligns with the U.S. Supreme Court's narrow judgment in the case. The estimated effect of the

234. *See* Edwards, *Survey Results*, *supra* note 63, at *Hopkins survey data.tab*. The estimated jury pool verdict preferences in the *Hopkins* study translate to P(G | actual) and P(G | hypo) values of .975 and .985. *Id.* The defendant will almost certainly be found guilty with or without the evidence of his confession. The margin of error of the harm estimate in *Hopkins* is ±.012, which is relatively small. To some extent, this reflects its larger sample size, but it is primarily due to the non-linear relationship between jury pool preferences and verdict probabilities. Figure 7 shows how the margin of error of P(g) relates to the margin of error of P(G). *See supra* Figure 7. When the P(g) values are very high or very low, the curve is relatively flat and uncertainty about P(g) does not create much uncertainty about P(G).

235. Some behavioral economics research suggests that low-quality evidence may reduce jurors' assessment of the overall strength of the prosecution's case. Economists have found that adding low-value objects to a set of goods lowers prospective buyers' overall evaluation of the set, even though objects with value have been added. *See* Christopher K. Hsee, *Less is Better: When Low-Value Options are Valued More Highly than High-Value Options*, 11 J. BEHAV. DECISION MAKING 107, 117 (1998) (adding broken items to dinnerware set lowers its market value); John A. List, *Preference Reversals of a Different Kind: The "More is Less" Phenomenon*, 92 AM. ECON. REV. 1636, 1639 (2002) (adding modest value baseball cards to a set of high value cards lowers the sale price of the entire set). In the *Hopkins* study, the involuntary confession casts some doubt on the reliability of forensic evidence. In the words of one survey respondent: "If the police are capable of coercion in the interrogating room, they are also capable of messing up the actual samples of DNA at the crime scene." Edwards, *Survey Results*, supra note 63, at *Hopkins survey data.tab*.

236. *See* Hopkins v. Cockrell, 325 F.3d 579, 586 (5th Cir. 2003).

237. Fulminante's involuntary confession increases the probability of conviction by an estimated .086, but given the inherent uncertainty of sample statistics, it is plausible that the effect was as high as .159, which is more than the suggested .100 harm tolerance. *See* Edwards, *Survey Results*, *supra* note 63, at *Fulminante survey data.tab*; *see also supra* Table 1 (providing verdict preferences); *supra* Part II (providing methodology); *supra* note 231 and associated text.

coerced confession is a mild increase in the probability of conviction, but the uncertainty of estimation with small samples precludes proof of harmlessness. The *Fulminante* study aligns with the Court's decision, but if the sample sizes were larger, the analysis may prove that the dissenting justices were correct.[238]

C. Improper Arguments by Prosecutors

This Part analyzes the fairness of two trials in which prosecutors made improper statements during their closing arguments in murder trials. The question in both appeals was whether the state could prove that the improper prosecutorial statements were harmless errors. The U.S. Supreme Court sided with one defendant, while the U.S. Court of Appeals for the Third Circuit ruled against the other defendant. Following the method outlined in Part II, we produce consistent results through quantitative analysis of juror verdict preferences.

The prosecutor in Ruth Chapman's criminal trial repeatedly commented on the defendant's decision not to testify.[239] The U.S. Supreme Court held that the prosecutor's comments violated Chapman's constitutional right to remain silent at trial but determined that the violation did not require automatic reversal.[240] The Court held that the State could not prove that the prosecutor's statements against Chapman were harmless.[241]

In *Lee v. Smeal*, the Third Circuit analyzed a prosecutor's improper references to a co-defendant's incriminating statements during the closing arguments of a first-degree murder trial in Pennsylvania.[242] The issue arose in post-conviction proceedings, where the defendant bore the burden of proving

238. The difference of opinion in *Fulminante* may result from the Justices focusing on different types of harm. A close reading of Justice Kennedy's analysis suggests he focused on the confession's effect on jurors rather than the verdict. Justice Kennedy does not explicitly identify his preferred approach to harmless error analysis, but he discusses the error's effect on the jury rather than its effect on the defendant. Arizona v. Fulminante, 499 U.S. 279, 313 (1991) (Kennedy, J., concurring) (discussing the "indelible impact a full confession may have *on the trier of fact*" and whether a jury "will be tempted to rest its decision on that evidence alone, without careful consideration of the other evidence in the case" (emphasis added)). Chief Justice Rehnquist, writing for the Court, also does not explicitly identify his approach to harmless error analysis. However, he suggests an effect-on-defendant approach. *Id.* at 312 (majority opinion) (acknowledging that an involuntary confession "in particular cases . . . may be devastating *to a defendant*" but concluding that the error was harmless in Fulminante's trial (emphasis added)).

239. Ruth Chapman was convicted of murder, robbery, and kidnapping in 1963. People v. Teale, 404 P.2d 209, 214 (1965), *rev'd sub nom.*, Chapman v. California, 386 U.S. 18, 18–19 (1967). At the time of her trial, the California State Constitution permitted prosecutors to comment on, and juries to draw negative inferences from, a defendant's failure to testify. *Chapman*, 386 U.S. at 19. The U.S. Supreme Court's opinion in the case does not offer a detailed summary of Chapman's crimes and subsequent trial, but additional facts can be found in *Teale*, 404 P.2d at 212–14. At the time, constitutional trial errors such as these prosecutorial arguments were thought to affect the criminal defendant's "substantial rights" and therefore required automatic reversal. *See* 28 U.S.C. § 111 (mandating harmless error review for errors and defects "which do not affect the substantial rights of the parties").

240. *Chapman*, 386 U.S. at 21–22.

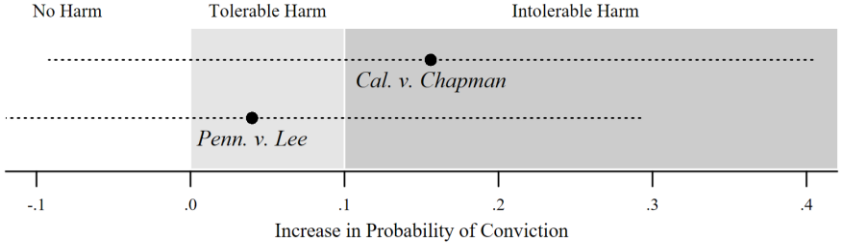
241. *Id.* at 24–26.

242. Lee v. Smeal, 447 F. App'x 357, 359 (3d Cir. 2011).

the trial error was harmful.[243] The Third Circuit held that the prosecutor's closing statements violated the defendant's right to confront witnesses, but that this error was not clearly harmful because there was other evidence of the defendant's guilt.[244]

Consistent with the discussion in Part II.A, the researchers randomly assigned online respondents to read trial summaries with or without the prosecutor's improper statements and then decide whether the defendant should be found guilty or not guilty.[245] Translating the *Chapman* study's P(g | actual) and P(g | hypo) estimates of verdict probabilities in California capital felony trials,[246] it is estimated that the prosecutor's comments on Chapman's refusal to testify increased the probability of conviction by .156.[247] With jury size and peremptory strike availability adjusted according to Pennsylvania law,[248] it is estimated that improper prosecutorial arguments caused a .040 increase in the probability of conviction in the *Lee* case.[249] Both harm estimates are presented in Figure 10, along with their margins of error. The small samples used in the *Chapman* and *Lee* studies result in relatively large margins of error.

Figure 10. Effect of Improper Prosecutorial Statements on Probability of Conviction



In *California v. Chapman*, most of the confidence interval for the harm estimate exceeds the threshold of tolerable harm. Based on this analysis, which aligns with the U.S. Supreme Court's decision, the state does not meet its burden of proving the error was harmless.[250]

---

243. *Id.* at 361 (citing Brecht v. Abrahamson, 507 U.S. 619, 637 (1993)).

244. *Id.* at 360–62.

245. *See supra* Table 1.

246. CAL. CIV. PROC. CODE § 220 (West 2025) (twelve-person jury rule); *id.* § 231(a) (both sides entitled to ten peremptory strikes).

247. In the *Chapman* study, the probability of a guilty verdict under actual trial conditions, P(G | actual), is .576, while under hypothetical trial conditions, P(G | hypo) is .420. *See supra* Table 1 (providing verdict preferences); *supra* Part II (providing methodology).

248. Pennsylvania law requires juries of twelve in felony trials. *Cf.* 234 PA. CODE § 6.641 (2025) (requiring party consent to have a jury of less than twelve). In capital felony prosecutions, the defendant is entitled to twenty peremptory challenges, as is the prosecution. 234 PA. CODE. § 6.634(A)(3) (2025)

249. The probability of a guilty verdict in the actual trial condition, P(G | actual), is .784, whereas the probability of a guilty verdict in the hypothetical trial condition, P(G | hypo), is .744. *See supra* Table 1 (providing verdict preferences); *supra* Part II (providing methodology).

250. The confidence interval is relatively wide due to the small sample size. However, a larger sample is unlikely to yield a more favorable result for the State in this case. The estimated harm, .156, exceeds the

While the prosecutor's improper comments in *Lee* may have, in reality, caused intolerable harm, the analysis does not demonstrate this to a reasonable degree of certainty. Based on these results, petitioner Lee would not carry his burden of proving that the trial error was harmful. This result is consistent with the Third Circuit's judgment.[251]

D. Effect of Jury Instructions

The remainder of Part III analyzes the results of two studies about the effect of jury instructions on jurors' decisions about guilt and innocence in the context of real criminal trials. Unlike the studies discussed in earlier parts, these were not cases where appellate courts evaluated trial errors. Nevertheless, they are excellent studies that show the method's versatility.

This Article discussed and referenced Simon's classic study of the effect of jury instructions in *U.S. v. King* throughout Part II to illustrate how fairness can be measured. In the *Berkowitz* case, the victim did not consent to intercourse and repeatedly said "no," but the defendant did not use force or the threat of force to compel the victim to engage in intercourse.[252] While not a study of trial error per se, it contributes to the debate over the legal definition of rape.[253]

Figure 11 presents the estimated effects of varying jury instructions on conviction probabilities in the *Berkowitz* and *King* trials. In *Berkowitz*, following Pennsylvania rules for jury size and peremptory strikes,[254] instructing jurors that "no means no" increases the probability of conviction by .257 compared to the common law definition of rape.[255] In the *King* case, based on twelve-person

asserted limit of tolerable harm of .10, suggesting significant harm by the prosecutor's inappropriate comments.

251. The *Chapman* and *Lee* studies do not support the claims made by the parties bearing the burden of proof. *See supra* Figure 10. This is partly due to the small sample sizes not yielding precise estimates. However, assuming that larger samples produce similar responses, the *Chapman* study would suggest the trial error caused intolerable harm, disproving the state's claim that the error was harmless, while the *Lee* study would indicate the trial error's harm was tolerable, disproving the petitioner's claim that the error was harmful. *Id.*

252. *See* Kahan, *supra* note 65, at 736–40.

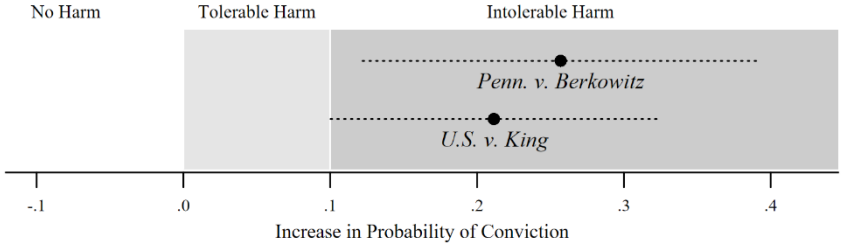
253. At the time of the *Berkowitz* trial, most states, including Pennsylvania, regarded force or threat of force as an element of the crime of rape. *See id.* at 739, 745. As Kahan notes, "the debate among legal commentators, at least, is less about whether *Berkowitz* was correctly decided than about whether the standard definition of rape that it enforced should be changed." *Id.* at 745.

254. Pennsylvania requires twelve-person juries in felony trials unless the parties and judge agree to a smaller jury. *See* 234 PA. CODE § 6.631 (2025) (procedures for impaneling twelve jurors); *id.* § 6.641 (agreement to less than twelve jurors). In non-capital felony cases, both the state and defendant are allowed seven peremptory challenges. *Id.* § 6.634(A)(2).

255. Kahan estimates a .120 difference in individual-level verdict preferences, but the difference in verdict probabilities is more than twice as large. *See generally* Kahan, *supra* note 65, at 773–791 (discussing the results of his statistical analysis on juror preferences); *supra* Part II (providing methodology for estimating verdict probabilities). This happens because the proportion of jurors who support a guilty verdict falls in the range where the S-curve has the steepest slope. *See supra* Figure 5 (showing the S-curve of individual juror preference to verdict probability). Most juries would be deeply divided between a guilty and not guilty verdict, so a change in one juror's vote can significantly change the probability of a guilty verdict.

juries with each side entitled to ten peremptory strikes,[256] the effect of varying the definition of insanity in jury instructions is a .211 increase in the probability of a guilty verdict.[257]

Figure 11. Effect of Varying Jury Instructions on Probability of Conviction



Would a mistaken jury instruction on the definition of rape in *Berkowitz* constitute harmful error? What if, for example, the trial judge instructed the jury that "no means no" while the state adhered to the common law definition?[258] Referring to Figure 11, the confidence interval for the effect of varying jury instructions in *Berkowitz* is relatively large, but the lower bound of the estimate (.122) still represents an intolerable level of harm. Therefore, a mistaken jury instruction would cause intolerable harm to the defendant.

In the *King* case, the lowest plausible effect of a mistaken definition of insanity is a .1002 increase in the probability of conviction.[259] If one assumes that a .10 increase in P(G) due to trial error is tolerable,[260] this analysis of *King* shows that a mistaken instruction on the definition of insanity constitutes harmful error to a reasonable degree of certainty.[261]

---

256. In the District of Columbia, criminal trial juries have twelve members unless the parties and court agree to a lesser number. D.C. CODE § 16-705(c) (2025). In capital trials, each side is allowed twenty peremptory challenges; in felony trials, each side is allowed ten strikes; in all other criminal trials, each side is entitled to three strikes. *Id.* § 23-105(a).

257. The probability of a guilty verdict in the actual trial condition, P(G | actual), is .875., whereas the probability of a guilty verdict in the hypothetical trial condition, P(G | hypo), is .663. *See supra* Table 1 (providing verdict preferences); *supra* Part II (providing methodology).

258. I am not arguing that this instruction is a trial error. State laws vary, but the defendant's only viable error claim arises when a judge mistakenly defines rape too broadly.

259. *See supra* notes 171–72 and accompanying text. Simon does not argue that it is an error to instruct jurors based on the *M'Naghten* rule rather than the *Durham* rule. *See* SIMON, *supra* note 56. Both definitions of insanity have been used in American courts. *See* 41 AM. JUR. 2D *Proof of Facts* 615 §§ 3, 6 (2025).

260. *See A Scientific Framework*, *supra* note 13, at 43–45.

261. The lower bound of the probability of a different outcome following the *M'Naghten* rule rather than the *Durham* rule in the *King* trial (.1002) is slightly more than the presumed tolerance value of .10. *See supra* Table 1 (providing verdict preferences); *supra* Part II (providing methodology). The analysis shows intolerable harm to a reasonable degree of certainty. This shows why a formal model of deliberation is useful. If we rely on Part II.C.4's empirical analysis, additional uncertainty about the relationship between juror preferences and jury verdicts enlarges the plausible range of the effect of mistaken instructions to .094 on the low end and .318 on the high end. A mistaken instruction on the definition of insanity is not proven harmful to a reasonable degree of certainty. Nor is a mistaken instruction proven harmless to a reasonable degree of certainty.

This Part applied the method of measuring fairness discussed in this Article to a variety of real criminal trials. It was used to assess the effects of ineffective defense counsel, coerced confessions, improper arguments by prosecutors, and varying jury instructions. The method was applied to wrongfully omitted evidence, mistakenly admitted evidence, post-conviction proceedings, and direct appeals.[262] The method of measuring fairness discussed in this Article accurately replicated the ultimate judgment of appellate courts in all cases where such comparisons were possible.[263] Additionally, the analysis supports studies by Kahan and Simon on the impact of jury instructions in criminal trials.

IV. DIRECTIONS FOR FUTURE RESEARCH

This Part explores extensions and additional implications of this research. While the model of jury deliberation developed here addresses a variety of criminal trial outcomes, civil jury trials remain a promising area for further research. Understanding the relationship between jurors' initial verdict preferences and jury verdicts in civil jury trials would aid in evaluating the effect of trial errors in civil appeals.[264]

In this case, the formal model's precision benefits the defendant, allowing him to demonstrate that a mistaken insanity instruction is harmful rather than leaving the analysis inconclusive. Precision happens to benefit the defendant in this example, but additional precision also benefits prosecutions when they aim to prove an error was harmless. For example, in the *Fulminante* and *Lee* cases, discussed above, analysis suggests tolerable error. *See* discussion *supra* Parts III.B.–C. However, estimation uncertainty leaves one unable to rule out the possibility of intolerable harm in these cases; greater precision would likely benefit the prosecution's position in these cases.

262. Although none of the examples from real trials involved *Brady* violations or actual innocence claims based on the discovery of new evidence, the method described in this Article is flexible enough to estimate the probability of different outcomes in these situations too. *See supra* Parts II.C–D.

263. Robertson studied one additional case: *State v. Jennings*, 716 S.E.2d 91 (S.C. 2011). In *Jennings*, the South Carolina Supreme Court held that the improper admission of expert testimony on the truthfulness of victims was not proven harmless. *See id.* at 93. Translating the study's individual-level results to guilty verdict probabilities, improper expert testimony increased P(G) by .286 ±.176. *See* Winkelman et al., *supra* note 64, at 1424 (providing P(g | hypo)=0.22, P(g | actual)=0.50, and, splitting the sample size in half, n=76); S.C. CODE ANN. § 14-7-1110 (providing for ten strikes for defendants and five for the prosecution). The *Jennings* trial error created a significant probability of a different outcome. *See id.* The margin of error is relatively large, as P(g) values were estimated with only seventy-six responses each. *See id.* Nevertheless, the results comport with the South Carolina Supreme Court's analysis of the trial error.

The author has recently completed additional applied research to estimate the effect of trial errors and omissions on the probability of death sentences to assist habeas corpus litigation. The author applied the methods discussed in this Article to analyze how a coparticipant's false testimony that he did not receive a lesser sentence for testifying affected the probability of a death sentence. *See* Brief of Amicus Curiae Fair Trial Analysis, LLC in Support of Warren King's Petition for Writ of Habeas Corpus at 5, King v. Emmons, No. 2024-SU-HC-0010 (Ga. Super. Ct. Feb. 5, 2025). The author also applied the methods of analysis discussed here to estimate the effect of prejudicial evidence in the guilt and sentencing phases of a capital trial. *See* Brief of Amicus Curiae Fair Trial Analysis, LLC in Partial Support of Petitioner and Partial Support of Respondent at 3–4, Andrew v. Tinsley, 164 F.4th 789 (10th Cir. 2026) (No. 15-6190). The analysis of *Okla. v. Andrew* addresses the questions that the U.S. Supreme Court directed the Tenth Circuit Court of Appeals to answer on remand. *See* Andrew v. White, 604 U.S. 86, 96 (2025).

264. *See* McDonough Power Equip., Inc. v. Greenwood, 464 U.S. 548, 553–54 (1984) (noting that civil appeals are not overturned due to legal technicalities and that the harmless error rule is codified by Federal

There are several reasons to believe that the relationship between juror preferences and jury verdicts differs between civil and criminal cases. First, the burden of proof is lower in civil cases than it is in criminal cases, and the defendant is not afforded a favorable presumption.[265] The plaintiff is not required to prove the cause of action beyond a reasonable doubt in civil cases.[266] For these reasons, the curve representing the relationship between juror preferences and jury decisions may be symmetrical in civil trials.[267] However, many different curves can be symmetrical. The shape of the curve depends on how responsive civil juries are to majority factions.[268]

Unfortunately, the relationship between initial polls and jury verdicts in civil matters cannot be evaluated with observational data as was done with criminal trials.[269] Research on the relationship between the initial preferences of jurors and jury decisions in civil trials is limited.[270] Further research is also needed to understand how juries decide interval-level outcomes, such as monetary damage awards.

Thus far, this Article has focused on trials where the outcome is a single variable, such as the decision whether to convict or to impose a death sentence. The relationship between jurors' initial preferences and verdicts becomes more complex when jurors aggregate preferences on multiple dimensions, which

---

Rules of Civil Procedure and federal statutes); FED. R. CIV. P. 61 ("At every stage of the proceeding, the court must disregard all errors and defects that do not affect any party's substantial rights."); *see also* 28 U.S.C. § 2111 (similar).

265. *See* Addington v. Texas, 441 U.S. 418, 423–24 (1979); Lilienthal's Tobacco v. United States, 97 U.S. 237, 267 (1878); *see also* J. Harvie Wilkinson III, *The Presumption of Civil Innocence*, 104 VA. L. REV. 589, 612 (2018) ("[T]here is no comparable recognition of a presumption of *civil* innocence.").

266. *See Addington*, 441 U.S. at 423–24.

267. This symmetry may be limited to cases where jurors decide a single issue. The relationship between initial preferences and verdicts when jurors decide multiple issues is less clear. *See* discussion *infra* Part IV.

268. Research suggests that the probability that the jury sides with the plaintiff is equal to the proportion of jurors initially favoring a plaintiff's verdict. *See A Study of Juror and Jury Judgments in Civil Cases*, *supra* note 113, at 305. However, when civil juries decide whether to impose punitive damages on a defendant, there may be a pro-defendant leniency shift. *See id.* at 305–06.

269. *See* Part II.C.4.

270. *See generally* James H. Davis, *Group Decision and Social Interaction: A Theory of Social Decision Schemes*, 80 PSYCH. REV. 97 (1973) (proposing a general theory of group decision-making that explains how social interaction changes individual preferences into collective decisions); Garold Stasser, Norbert L. Kerr & James H. Davis, *Influence Processes and Consensus Models in Decision-Making Groups*, *in* PSYCHOLOGY OF GROUP INFLUENCE 279 (Paul B. Paulus ed., 2d ed. 1989) (illustrating the efficacy of computer models to preduct possible group decisions and reflecting on their application); ROBERT E. LITAN ET AL., VERDICT: ASSESSING THE CIVIL JURY SYSTEM (Robert E. Litan ed., 1993) (compiling fifteen essays addressing the history, performance, and value of juries in civil litigation); Robert MacCoun, *Inside the Black Box: What Empirical Research Tells Us About Decisionmaking by Civil Juries*, *in* VERDICT: ASSESSING THE CIVIL JURY SYSTEM 137 (Robert E. Litan ed., 1993) (noting the anecdotal nature of arguments against the value of civil juries and arguing that the civil-jury system is not inherently flawed); Michael J. Saks et al., *Reducing Variability in Civil Jury Awards*, 21 L. & HUM. BEHAV. 243 (1997) (testing and comparing the efficacy of procedures for reducing variability in civil damage awards); Paula Hannaford-Agor et al., *The Timing of Opinion Formation by Jurors in Civil Cases: An Empirical Examination*, 67 TENN. L. REV. 627 (2000) (conducting a retrospective empirical analysis to identify when jurors form their opinions during civil cases); H.P. Weld & E.R. Danzig, *A Study of the Way in Which a Verdict Is Reached by a Jury*, 53 AM. J. PSYCH. 518 (1940) (describing the effect of discussion in the jury room on jurors' decisions).

occurs in civil cases when jurors deliberate liability and damages simultaneously and in criminal cases with alternative verdicts. When jurors must reach agreement on two dimensions—liability and damages—they may compromise between the two in order to reach a consensus.[271] If, for example, one faction prefers a large award for the plaintiff while the other prefers a defense verdict (and thus no damages), the jury may compromise by awarding a smaller award to the plaintiff.[272] With initial preferences distributed across two dimensions, a range of verdicts is possible, depending on how jurors deliberate and their willingness to compromise.[273] When groups aggregate individual preferences along two dimensions, the relationship between individual preferences and group decisions becomes notoriously difficult to identify.[274] In this scenario, jurors' initial preferences may translate into a set of plausible outcomes, representing combinations of verdict probabilities and dollar amounts for damages.

Similar complications arise in criminal cases when jurors have several verdict options. The elements of some crimes encompass lesser offenses; a jury may find a defendant guilty of a lesser offense than the one charged.[275] A defendant can also be charged with multiple crimes that do not share common elements.[276] When verdict possibilities encompass lesser included offenses, multiple counts, or multiple charges, the factions competing in deliberation are less well-defined. When individuals can rank multiple alternatives, no social choice function will consistently yield stable group decisions.[277] Unstable majorities and voting cycles can emerge from pairwise comparisons of charges and verdict options.

---

271. This sort of compromise is not possible in criminal cases where the jury determines guilt and the judge decides the length of the sentence. In capital cases, the guilt and punishment phases are intentionally bifurcated to prevent compromise verdicts. *See* Morris B. Hoffman, *The Case for Jury Sentencing*, 52 DUKE L.J. 951, 989, 1004–05 (2003).

272. Jurors are not supposed to compromise for the sake of expediency, but they appear to do so anyway. *See generally* Michael D. Craig, Note, *Improving Jury Deliberations: A Reconsideration of Lesser Included Offense Instructions*, 16 U. MICH. J.L. REFORM 561 (1983) (advocating for lesser included offense instructions that are flexible without undermining the deliberative process); Samuel Bray, Comment, *Not Proven: Introducing a Third Verdict*, 72 U. CHI. L. REV. 1299 (2005) (proposing a third verdict: "not proven").

273. In this situation, the jurors' initial preferences can be mapped in a two-dimensional space with the liability decision on one axis and damages preferences along the other.

274. *See* Richard D. McKelvey, *Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control*, 12 J. ECON. THEORY 472, 472 (1976) ("The implications of this result are that it is theoretically possible to design voting procedures which, starting from any given point, will end up at any other point in the space of alternatives, even at Pareto dominated ones."). *See generally* William H. Riker, *Implications from the Disequilibrium of Majority Rule for the Study of Institutions*, 74 AM. POL. SCI. REV. 432 (1980) (noting this difficulty and recommending that political scientists pursue research into institutions instead of tastes).

275. FED. R. CRIM. P. 31(c). For example, a defendant on trial for first-degree murder can be found guilty of second- or third-degree murder, manslaughter, battery, assault, or attempted crimes.

276. *See* FED. R. CRIM. P. 8(a) (joinder of offenses).

277. *See generally* KENNETH J. ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES (1951) (analyzing the limits of aggregating individual preferences into collective decision-making); ERIC MASKIN ET AL., THE ARROW IMPOSSIBILITY THEOREM (2014) (remarking on the legacy of Arrow's work and suggesting avenues for exploration within the field of social choice).

Further research may shed light on how juries deliberate when there are more than two possible verdicts. If the verdict options can be ordered, as is the case with lesser included offenses, Pennington and Hastie suggest that the deliberation process can be understood as deliberation with multiple dividing lines.[278] This article has not addressed multi-dimensional deliberations in civil or criminal trials, but it remains an intriguing and important area for further research.

CONCLUSION

This Article addresses a fundamental question in criminal law: Did the defendant receive a fair trial? Although this inquiry is central to criminal justice, appellate courts lack a systematic approach to evaluate whether a trial error or omission created a "reasonable probability of a different outcome."[279] Even when a trial error or omission is well-defined, existing methods do not permit its effect to be "quantitatively assessed in the context of other evidence presented," rendering the right to a fair trial an unenforceable promise.[280]

Measuring fairness requires bridging the gap between individual juror preferences and the probability of a guilty verdict from jury deliberation. This Article introduces a method of linking initial juror preferences to overall verdict probabilities, providing a structured approach to evaluating the probability of a different outcome in the absence of a trial error or omission. By effectively contextualizing survey data, this method enables researchers to extrapolate verdict probabilities from juror preferences. This approach allows researchers to estimate jury pool preferences, obtain corresponding verdict probabilities, and quantify the extent to which a trial error or omission altered the probability of conviction.

This Article demonstrates the validity and reliability of the measure on multiple levels. The deliberation model linking preferences to verdict probabilities accurately represents how juries deliberate.[281] The method replicates appellate courts' independent analyses of trial errors and omissions in a series of test cases concerning ineffective assistance of counsel, coerced confessions, and improper prosecutorial arguments.[282] This effort to ensure fair trials would benefit from additional validation studies of criminal trials, the

278. *See* HASTIE, PENROD & PENNINGTON, *supra* note 107, at 185–95. Deliberation over ordered verdict options may be an extension of the Markov chain model's deliberation between two verdict options in Part II.C.2.

279. Weaver v. Massachusetts, 582 U.S. 286, 303 (2017) (citing Strickland v. Washington, 466 U.S. 668, 694 (1984)). The "reasonable probability of a different outcome" standard governs a variety of claims, including direct appeals arising out of trial errors, petitions for post-conviction relief, and motions for new trials based on *Brady* violations or newly discovered evidence. *See Strickland*, 466 U.S. at 694.

280. Arizona v. Fulminante, 499 U.S. 279, 307–08 (1991).

281. *See* discussion *supra* Part II.C.4.

282. *See* discussion *supra* Part III.

development of standard protocols for survey-based estimates of verdict preferences in jury pools, and future research on jury deliberation in civil trials and cases with multiple potential outcomes.

This research seeks to improve criminal justice processes by offering a scientifically grounded framework for assessing trial fairness. While this research holds promise, there are reasons for tempered expectations. Historically, courts have been reluctant to acknowledge fallibility, even when confronted with seemingly irrefutable evidence of wrongful convictions.[283] But there are also reasons to be cautiously optimistic. One reason that appellate courts struggle with claims of actual innocence is that they are better equipped to assess fairness and address trial errors. A fairness measure aligns with the traditional role of appellate courts, supporting, rather than challenging, judicial authority.[284] Researchers can help courts carry out the difficult and essential task of judging whether defendants received fair trials. Judges, in turn, should recognize the limits of their ability to estimate the probability of a different trial outcome.[285] Understanding how juries are likely to decide specific cases is challenging without conducting systematic research.[286] Given that courts have gradually accepted scientific research—including survey data—in other contexts,[287] one hopes that they will accept assistance in quantifying the probability of different trial outcomes.

There are additional reasons for cautious optimism. The fairness measure described in this Article requires little more than the willingness to use research to better understand jurors and juries. It requires no new laws, changed legal standards, or investment of public resources. The necessary case-specific data can be efficiently obtained using existing survey research methods and technologies. To ensure accessibility, the deliberation model and analytical functions discussed in this Article are freely available in an open-source

283. *See* ROBERT J. NORRIS, EXONERATED: A HISTORY OF THE INNOCENCE MOVEMENT 25–29 (2017) (discussing the persistence of the "myth of systemic infallibility"); Garrett, *supra* note 11, at 117–20 (describing how courts limited access to DNA evidence and delayed or refused the release of those exonerated). Courts were not eager to free innocent men and women through post-conviction remedies; they were forced to give access to DNA evidence and allow these claims by federal and state laws. *See id.* at 120 (noting that only two of the first 200 individuals exonerated by DNA evidence were freed through federal habeas corpus proceedings).

284. The method described in this Article relies on, and cannot replace, legal and normative judgements made by appellate courts. An estimate of a confession's effect, for example, does not determine whether the confession was involuntary. Estimates must be interpreted in the appropriate procedural context to know which party bears the burden of proof and how much certainty is required to satisfy the burden of proof. Additionally, determining how much unfairness we are willing to tolerate—the dividing line between fair and unfair—requires normative judgment instead of empirical analysis.

285. Judges may believe they can rationally assess the fairness of trials, but they must recognize the inherent limitations of their ability to predict jury behavior. Judges should not assume that they know the probabilities of different outcomes, as no court has attempted to estimate them, or even discuss how they might be quantified. The standard asks judges to do something they are not equipped or able to do.

286. The researcher cannot estimate the harmfulness of a trial error or omission without studying sample data.

287. *See supra* notes 54–56 and accompanying text.

statistical software package for those who may benefit from an objective measure of fairness.[288] By comparing two verdict probabilities, as if weighing trials on the scales of justice, we can determine the probability of a different outcome and objectively assess whether a defendant's trial was fair.

288. *See SATE*, *supra* note 12.