AN EMPIRICAL METHOD FOR HARMLESS ERROR

D. Alex Winkelman, David V. Yokum, Lisette C. Cole, Shelby C. Thompson, Christopher T. Robertson*

ABSTRACT

Trials are often imperfect. When inadmissible evidence is introduced or the jury is incorrectly instructed, judges must determine whether the error was prejudicial or merely harmless. In making that assessment, judges resort to speculation about the counterfactual question of whether the error changed the outcome, compared to the decision of a properly informed and instructed jury. These decisions are likely colored by confirmation and status quo biases, along with "mental contamination" of the error itself. Even when appellate judges perform these analyses accurately, their decisions appear conclusory. Scholars and judges have roundly criticized this doctrine, but no solution has emerged.

We developed and piloted an unbiased and transparent method for making harmless error determinations, using randomized experiments with simulated jurors. To pilot this method on three real cases, we recruited 489 human subjects to participate as mock jurors reviewing trial vignettes that we manipulated into conditions with and without the errors. Subjects were blinded to the purpose of the study and to the first trial's outcome. By comparing verdict rates in the error and no-error conditions, we estimated whether the error was harmful.

We found a high degree of correspondence between the assessments of real judges and our experimental method, which could be taken as a validation of the method and reassurance that it would not cause a radical change in the rates at which new trials are granted. Still, across the thousands of cases in which harmless error determinations are made each year, the empirical method may be more reliable since it avoids known biases. The transparency of our method may also lend greater legitimacy to harmlessness determinations. If such a method is used as a tool for litigants in real cases,

---

*. Winkelman, Yokum, Cole, and Thompson were students at the University of Arizona at the time this study was conducted. Robertson is visiting professor, Harvard Law School, and associate professor, James E. Rogers College of Law, University of Arizona. He may be contacted at crobertson@post.harvard.edu. The authors thank the Honors College at the University of Arizona for funding these experiments, and thank Harvard Law School JD student Jaime McFarlin for excellent research assistance.

courts will be called upon to establish procedures for taking such evidence and then draw lines to specify how much prejudice is too much, while also being sensitive to the limitations of statistical power. Our study is most useful as proof of concept for a new method to improve harmless error analyses.

TABLE OF CONTENTS

I. THE DIFFICULTY OF DETERMINING HARMLESSNESS

At the turn of the twentieth century, appellate courts were reversing criminal convictions at such a rate that these courts were called "impregnable citadels of technicality."[1] In response to public outcry, Congress passed a statute in 1919 which "declared that convictions shall not be reversed for 'errors or defects which do not affect the substantial rights of the parties.'"[2] This concept has become known as the harmless error doctrine.[3] It is now one of the most consequential doctrines in jurisprudence; in almost *every* case where an error is found, the court must also determine whether it was harmless.[4] Strikingly, in some cases the harmless error determination is so important that it averts the need for the court to even consider the merits of the underlying error alleged.[5] A similar concept of "actual prejudice" applies to post-conviction review,[6] and courts also apply the harmless error doctrine in civil contexts.[7]

The exact legal analysis that judges perform, and should perform, is the matter of some debate.[8] In function, however, the harmless error doctrine allows appellate judges to affirm criminal convictions and other appealable judgments even though the trial was infected with error; the doctrine averts the need for a second error-free trial before a jury, which would otherwise be required by the Constitution.[9] This doctrine is applicable even to trial errors that infringe constitutionally protected rights, such as the Fifth Amendment

---

1. Charles S. Chapel, *The Irony of Harmless Error*, 51 OKLA. L. REV. 501, 522 (1998) (citation omitted).

2. *Id.* (quoting 28 U.S.C. § 2111 (1994)).

3. *Id.*

4. Steven H. Goldberg, *Harmless Error: Constitutional Sneak Thief*, 71 J. CRIM. L. & CRIMINOLOGY 421, 421 n.3 (1980). Based on data from 1967 to 1980, the doctrine was pivotal in "possibly as high as ten percent of all criminal appeals." *Id.* at 421.

5. *See, e.g.*, Alzamora v. Sec'y Dep't of Corr., 463 F. App'x 816, 818 (11th Cir. 2012) ("We need not reach this question because we conclude that the failure to give this instruction, under the facts of this case, constituted harmless error . . . .")

6. *See* Murray v. Carrier, 477 U.S. 478, 485, 493 (1986) (balancing procedural default and prejudice against constitutional claims in the post-conviction setting).

7. David A. Shields, *East vs. West—Where Are Errors Harmless?*, 71 ST. LOUIS U. L.J. 1319, 1321 (2012) (noting that federal circuits have adopted standards for "non-constitutional error" in the civil setting).

8. *See* Chapel, *supra* note 1, at 503 ("The [Supreme] Court has been inconsistent in its analysis and rationale."); Martha A. Field, *Assessing the Harmlessness of Federal Constitutional Error*, 125 U. PA. L. REV. 15, 16 (1976). *See generally* Brandon L. Garrett, *Innocence, Harmless Error, and Federal Wrongful Conviction Law*, 2005 WIS. L. REV. 35, 36 (2005) (offering a critical view of the harmless error doctrine).

9. *See* U.S. CONST. amend. VI, § 1 (criminal jury trial right); U.S. CONST. amend. VII, § 1 (civil jury right); Field, *supra* note 8, at 16 (discussing the function of the doctrine).

right of self-incrimination.[10] The Supreme Court has said, however, that "structural" errors that infect the entire proceeding are not subject to harmless error review.[11]

In the federal system, the courts say that the harmless error doctrine allows them to affirm as long the error does "not affect the substantial rights of the parties."[12] Courts also ask "whether there is a reasonable possibility that the evidence complained of might have contributed to the conviction."[13] Only if the government has "failed to show overwhelming proper evidence of guilt or otherwise [that there was] no impact on the verdict (that is, failed to show that defendant suffered no prejudice to substantial rights), then the conviction is overturned."[14]

The harmless error doctrine was designed to preserve judicial resources and protect "public confidence in the judicial system," and scholars and judges have defended the doctrine as something of a regrettable necessity.[15] Others have criticized the doctrine as "one of the supreme ironies of our time,"[16] a "beast that swallowed the Constitution,"[17] and a "Constitutional

10. *See* Chapman v. California, 386 U.S. 18, 19–20 (1967); Field, *supra* note 8, at 25.

11. United States v. Gonzalez-Lopez, 548 U.S. 140, 148–49 (2006); Arizona v. Fulminante, 499 U.S. 279, 309 (1991). In *Arizona*, the Court gave "deprivation of the right to counsel" or denial of the right "to a judge who was not impartial" as examples of structural error. *Id.* For a deeper discussion of the effects of the trial and structural error distinction on our method, see *infra* Part V.

12. *Chapman*, 386 U.S. at 22 (citing 28 U.S.C. § 2111 (1967)); *see* Martha Davis, *Harmless Error in Federal Criminal and Habeas Jurisprudence: The Beast that Swallowed the Constitution*, 25 T. MARSHALL L. REV. 45, 55 (1999) (discussing the harmless error standard).

13. *Chapman*, 386 U.S. at 23 (citing Fahy v. Connecticut, 375 U.S. 85, 86–87 (1963)). To answer that question, courts can consider whether: erroneously admitted evidence might have contributed to a guilty verdict; whether once the erroneously admitted evidence is excluded, there remains overwhelming evidence to support the jury's verdict; or whether the tainted evidence is merely cumulative or duplicative of remaining evidence. Field, *supra* note 8, at 16.

14. Davis, *supra* note 12, at 55.

15. *See, e.g.*, Shannon L. Bybee, Jr., *A Comment on Application of the Harmless Constitutional Error Rule to "Confession" Cases*, 1968 UTAH L. REV. 144 (applying the harmless error rule to coerced confession cases); Kathleen M. Golden, *The Sequestration of Criminal Defendants: A Proposal for the Use of Harmless Error Analysis in the Aftermath of* Geders v. United States, 52 ALB. L. REV. 243, 246–47 (1987) (proposing harmless error analyses of erroneous sequestration orders); Steven K. Sharpe & John E. Fennelly, Massachusetts v. Sheppard*: When the Keeper Leads the Flock Astray—A Case of Good Faith or Harmless Error?*, 59 NOTRE DAME L. REV. 665, 666–67 (1984) (applying the concept to equitable suppression rulings); *The Harmless Error Rule Reviewed*, 47 COLUM. L. REV. 450 (1947) (arguing that it protects public confidence in the judicial system, which would otherwise erode due to frequent reversals); Sara E. Welch, *Supreme Court Review: Fifth Amendment—Harmless Error Analysis Applied to Coerced Confessions*, 82 J. CRIM. L. & CRIMINOLOGY 849, 849–50 (1992) (all errors should be subject to harmless error analysis).

16. Chapel, *supra* note 1, at 540.

17. Davis, *supra* note 12, at 45.

Sneak Thief."[18] The classic essay on the doctrine referred to it as a "riddle."[19] This sort of criticism is regrettably common.[20] As Charles Chapel has reviewed the literature: "The theme most common to the criticism is that the harmless error rule particularly offends the values attributed to our constitutional system of individual rights and liberties."[21]

A. Accuracy

Fundamentally, harmless error doctrine "assumes that an appellate judge in our system can in fact determine whether or not an error is harmless."[22] This inquiry is severely underdetermined by the information available to the decision maker. Some (but not all) appellate judges have experience as trial

---

18. Goldberg, *supra* note 4, at 421.

19. ROGER J. TRAYNOR, THE RIDDLE OF HARMLESS ERROR 33 (1970).

20. *See e.g.*, Kenneth R. Brown, *Constitutional Harmless Error or Appellate Arrogance*, 6 UTAH B. J. 18, 18 (1993) (doctrine promotes ends-justifies-the-means mentality); Steven D. DeBrota, *Arguments Appealing to Racial Prejudice: Uncertainty, Impartiality, and the Harmless Error Doctrine*, 64 IND. L.J. 375, 375–78 (1989) (should not apply to racially prejudicial arguments); Bennett L. Gershman, *The Gate Is Open but the Door Is Locked—Habeas Corpus and Harmless Erro*r, 51 WASH. & LEE L. REV. 115, 115–16 (1994) (the rule allows state officials to disregard constitutional norms); Goldberg, *supra* note 4, at 421 (doctrine is insidious because it destroys constitutional and institutional values); Craig Goldblatt, *Disentangling* Webb*: Governmental Intimidation of Defense Witnesses and Harmless Error Analysis*, 59 U. CHI. L. REV. 1239, 1239–41 (1992) (doctrine should not apply to governmental intimidation of defense witnesses); Jana J. Green, Arizona v. Fulminante*: The Harmful Extension of the Harmless Error Doctrine*, 17 OKLA. CITY U. L. REV. 755, 755–56 (1992) (doctrine should not apply to courts coerced confessions); Tamara Lynne Jones, *Coerced Confessions and Harmless Error*, 18 OHIO N.U. L. REV. 877, 877–78 (1992); Jason S. Marks, *Postscript: Harmless Error, Habeas Corpus, and a Constitutional Eclipse*, CRIM. JUST., Fall 1993, at 30, 30 (doctrine should not apply to constitutional violations); Gregory Mitchell, *Against "Overwhelming" Appellate Activism: Constraining Harmless Error Review*, 82 CALIF. L. REV. 1335, 1335 (1994) (variations in standard are problematic); Marla L. Mitchell, *The Wizardry of Harmless Error: Brain, Heart, Courage Required When Reviewing Capital Sentences*, 4 KAN. J.L. & PUB. POL'Y 51, 51–52 (1994) (the doctrine is inappropriate for capital cases); Henry P. Monaghan, *Harmless Error and the Valid Rule Requirement*, 1989 SUP. CT. REV. 195, 195 (doctrine allows judge-centered analysis, rather than jury-centered); Charles J. Ogletree, Jr., Arizona v. Fulminante*: The Harm of Applying Harmless Error to Coerced Confessions*, 105 HARV. L. REV. 152, 152–54 (1991); James C. Scoville, *Deadly Mistakes: Harmless Error in Capital Sentencing*, 54 U. CHI. L. REV. 740, 740–41 (1987) (doctrine should not apply to sentencing phase of a capital case); David M. Skoglind, *Harmless Constitutional Error: An Analysis of Its Current Application*, 33 BAYLOR L. REV. 961, 961 (1981) (benefits of doctrine fail to justify risks).

21. Chapel, *supra* note 1, at 505–06; *see e.g.*, Davis, *supra* note 12, at 48 (stating "what American jurisprudence has come to, as a result [of the harmless error doctrine], is a treatment of constitutional error, at least in criminal jurisprudence, as though the Constitution does not exist").

22. Chapel, *supra* note 1, at 516.

court judges, and perhaps this experience provides them with sufficient knowledge of how different sorts of evidence, argument, and instructions impact jury decisions. Nonetheless, some have suggested that the entire endeavor is futile: "a judge cannot possibly know or review what in the minds of the jurors led to the verdict."[23] Many have worried "that appeals court judges are ill-equipped by intuition to estimate the strength of the prosecutor's case and the cumulative or 'harmless' nature" of certain errors.[24]

There has been very little empirical investigation of this question. One recent study by Solomon used qualitative coding of 263 published habeas petitions from federal courts.[25] This study suggested, among other things, that "nearly two out of three analyses" resulted in a court finding that the error was harmless.[26] The study further suggested that harmless error is most commonly applied to improperly admitted evidence and errors in jury instruction.[27]

In Wallace and Kassin's vignette-based study of 132 judges making simulated harmless error determinations, 91% of the judges found that admission of a coerced confession would have been prejudicial (i.e., not harmless), but this finding was not sensitive to experimental manipulation of the strength of the other evidence.[28] Although four times as many judges would have themselves convicted in the experimental condition where the evidence was strongest, they found the error to be equally prejudicial in both the strong and weak conditions.[29] Although it is hard to say what the right answer is in the vignette provided, this finding suggests that judicial determination may not be sensitive to the facts of the case presented. Instead, judges may be performing some sort of categorical mode of reasoning, rather than actually considering the likelihood that the error changed the jury's decision.

---

23. *Id.*

24. D. Brian Wallace & Saul M. Kassin, *Harmless Error Analysis: How Do Judges Respond to Confession Errors*, 36 LAW & HUM. BEHAV. 151, 151 (2012).

25. James M. Solomon, *Causing Constitutional Harm: How Tort Law Can Help Determine Harmless Error in Criminal Trials*, 99 NW. U. L. REV. 1053, 1065 (2005). *See generally* Joep Sonnemans & Frans van Dijk, *Errors in Judicial Decisions: Experimental Results*, 28 J.L. ECON. & ORG. 687 (2012) (describing a laboratory experiment examining the relationship between evidence of which the diagnostic value is known, subjective probability of guilt, and errors in verdicts for abstract criminal cases, finding many mistakes, evenly divided over unfounded convictions and unfounded acquittals).

26. Solomon, *supra* note 25, at 1067.

27. *Id.* at 1066.

28. Wallace & Kassin, *supra* note 24, at 156.

29. *Id.* at 155–56.

B. Bias

Supposing that judges do have some prima facie ability to discern harmlessness, the skill is likely undermined by the suboptimal conditions in which appellate judges are called upon to render such a decision. Typically, judges are themselves exposed to the error (e.g., inadmissible evidence, or a prosecutor's comment on the defendant's failure to testify), and it likely shapes their own assessments of the underlying facts, and thus may tend to make their own view of the merits cohere with that of the trial jury, which was also exposed to the error.[30] Counterfactual reasoning is always difficult, but it is here profoundly hard to imagine what someone else would have decided if they did not know what the judge now knows. Psychologists call this exposure to irrelevant information a "mental contamination," and have shown that it is often nearly impossible to overcome such an exposure.[31] Indeed, psychological research has shown that individuals are simply unable to assess their own biases, which would be the first step towards correcting them.[32] Under a coherence-based reasoning model, the exposure of the improper evidence to the judge has the possibility of unconsciously shifting the judge's perception of the other pieces of evidence.[33] This is particularly worrying where the improper evidence is dispositive—or even merely highly suggestive—of guilt. Coherence theory would suggest that where judges have deemed an error harmless, they are cognitively predisposed to "dismiss[], reject[], or ignore[]" other aspects of the case that cut against their determinations.[34] This predisposition is particularly worrying where judges may merely announce that an error is in fact harmless—rather than clearly elucidate their reasoning.

In addition to the error, exposure to the outcome of the first trial is another contamination. In performing the harmless error analysis, judges often try to determine whether the case was a "close" one, which could easily be affected

---

30. *See* Alan Hirsch, *Confessions and Harmless Error: A New Argument for the Old Approach*, 12 BERKELEY J. CRIM. L. 1, 3 (2007); Wallace & Kassin, *supra* note 24, at 154 (judges exposed to high-pressure confession viewed guilt as more likely than those not so exposed).

31. *See* Timothy D. Wilson & Nancy Brekke, *Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations*, 116 PSYCHOL. BULL. 117, 117 (1994).

32. *See* Emily Pronin et al., *The Bias Blind Spot: Perceptions of Bias in Self Versus Others*, 28 PERS. SOC. PSYCHOL. BULL. 369, 378 (2002).

33. Under a coherence model, decision makers' perception will shift unbeknownst to them in order to make difficult mental tasks easier. Dan Simon, *A Third View of the Black Box: Cognitive Coherence in Legal Decision Making*, 71 U. CHI. L. REV. 511, 513 (2004). In particular, their perceptions of the other pieces of evidence will shift so that they cohere with outliers. *Id.* at 531.

34. *Id.* at 522.

by the error, or an easy case with many other bases for conviction aside from the error. In making this assessment, hindsight bias is relevant, as the judges try to conceive the defendant's risk of conviction ex ante, but have to make this estimate after the risk of conviction has actually materialized. Behavioral science, some performed on actual federal judges, suggests that close cases may begin to look like easy cases ex post, once the conviction has been rendered.[35]

The first trial's outcome also forms a new status quo.[36] For a criminal case, the appellant is no longer just a defendant; she is a felon. And, by the time a conviction reaches an appellate decision, a year or more may have passed, cementing the status quo, and increasing the mental switching cost.

There are also preferences and norm ascriptions at work. The appellate decision makers are invested in the very legal system and part of a social network with the trial court judges that committed the alleged errors. Thus knowledge of the trial outcome may also give rise to "confirmation biases," which tend towards affirming the trial court's result.[37]

For all these reasons, scholars have concluded that "the very enterprise of after-the-fact review is doomed to failure. Judges simply cannot see the errors because psychological biases make it hard to imagine that cases could have come out any differently."[38]

C. Legitimacy

Even supposing judges have this ability and can somehow overcome the obvious biases, there is a problem of appearances. In Solomon's textual analysis of the case law, he coded whether the judges were applying a particular test for harmlessness.[39] Solomon's data suggests not, finding that "[l]ess than 20% of the [judges'] analyses used a test for determining harm."[40]

---

35. *See* Chris Guthrie et al., *Inside the Judicial Mind*, 86 CORNELL L. REV. 777, 801–03 (2001) (reporting hindsight bias experiments with magistrate judges); Jeff Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 571, 572 (1998).

36. *See* PAUL BREST & LINDA HAMILTON KRIEGER, PROBLEM SOLVING, DECISION MAKING, AND PROFESSIONAL JUDGMENT: A GUIDE FOR LAWYERS AND POLICYMAKERS 423 (2010) ("*[S]tatus quo bias* [is] the general phenomenon whereby people attach a value to the present state of the world compared to alternative states.").

37. *Id.* at 609.

38. Stephanos Bibas, *The Psychology of Hindsight and After-the-Fact Review of Ineffective Assistance of Counsel*, 2004 UTAH L. REV. 1, 2; *see also* Keith A. Findley & Michael S. Scott, *The Multiple Dimensions of Tunnel Vision in Criminal Cases*, 2006 Wis. L. Rev. 291, 350–51 (discussing these problems).

39. Solomon, *supra* note 25, at 1067.

40. *Id.*

The remaining judges appeared to be making something like a gestalt, all-things-considered judgment call.

Whether they affirm or reverse, judicial determinations of the harmlessness of errors often read like arbitrary and conclusory pronouncements.[41] The inevitable dissents seem just as conclusory.[42] The nature of the task set before appellate judges—to determine whether a given error had an effect on a decision-making body that deliberates in secret—almost ensures a conclusory justification.

A recent example is illustrative. In 2011, a Seventh Circuit court considered whether the improper admission of evidence in the government's rebuttal constituted a harmless error.[43] In ruling that the error was harmless, the majority dove deeply into the facts for the charges concerning the defendant's flight from police:

> If there were degrees of flight, what happened here would be flight in the first degree. How else do you describe throwing the Bonneville into reverse, endangering officers (recall that Agent Chupik, with gun drawn, had to jump out of the way), hitting two police squad cars, and gunning it the wrong way into a roadway from the parking lot, ditching the car a few moments later and trying to escape by running through the kitchen and out the back door of a McDonald's?[44]

The dissenting judge, for his part, engaged in a similar assessment of the evidence, finding that the erroneously admitted piece of evidence "made for a fairly dramatic conclusion for the trial."[45]

---

41. *See, e.g.*, Chapman v. California, 386 U.S. 18, 25–26 (1967) ("[*Chapman*] was also a case in which, absent the constitutionally forbidden comments, honest, fair-minded jurors might very well have brought in not-guilty verdicts."); United States v. Runyon, 707 F.3d 475, 498 (4th Cir. 2013) ("What ultimately drove the jury's decision was not some video but the overpowering evidence of Runyon's guilt, his pivotal role in the crime, and the exceptionally callous nature of his conduct. With three fatal shots to the chest and abdomen, Runyon robbed an innocent man of his life and two small children of their father. And for what? Money."); Clemons v. State, 535 So.2d 1354, 1364 (Miss. 1988) ("We likewise are of the opinion beyond a reasonable doubt that the jury's verdict would have been the same with or without the 'especially heinous, atrocious or cruel' aggravating circumstance."), *vacated*, 494 U.S. 739 (1990); Satterwhite v. State, 726 S.W.2d 81, 93 (Tex. Crim. App. 1986) (after listing the evidence, the court concluded "that the properly admitted evidence was such that the minds of an average jury would have found the State's case sufficient on the issue"), *rev'd sub nom.* Satterwhite v. Texas, 486 U.S. 249 (1988).

42. *See, e.g.*, *Chapman*, 386 U.S. at 55 (Harlan, J. dissenting) ("The added impact of [the prosecutor's] comment would seem marginal in a case of this type . . . .").

43. United States v. Vasquez, 635 F.3d 889, 898 (7th Cir. 2011).

44. *Id.*

45. *Id.* at 899.

The Supreme Court has cautioned that an appellate judge must not "become in effect a second jury."[46] Cases like this make it difficult to see them as much else, especially where reasonable judges seem to differ.[47] The Supreme Court granted certiorari to address the question of harmlessness but then retreated, apparently feeling that it was unable to provide any better guidance or any practical test for harmlessness.[48]

It is striking to see a case that may have incurred millions of dollars of attorneys' time, and risks life or death for a litigant, all come down to the *ipse dixit* of a panel of judges concluding that admitted errors in the trial process seem harmless to them.[49] All the rest of due process can seem like cheap trappings for what is ultimately a conclusory disposition. In this sense, our contemporary method for applying harmless error doctrine has a fundamental problem of legitimacy.[50]

In this short Article, we develop a new method for investigating the harmlessness of errors, which could be used in real cases to inform judicial determinations. By using a blinded and randomized experiment, we can shed light on the counterfactual question of what the jury might have done if it had not been exposed to the error.

II. THE METHOD

A. Conception and Precedents

One must first clarify what exactly a judge is supposed to be assessing when performing a harmless error analysis. In accordance with other scholars, we suggest that the question is ultimately one of causation.[51] This

---

46. *Id.* at 901 (Hamilton, J., dissenting) (quoting Neder v. United States, 527 U.S. 1, 19 (1999)).

47. *See, e.g.*, Mest v. Cabot Corp., 449 F.3d 502, 516 (3d Cir. 2006) ("Generally, where reasonable minds can disagree, questions . . . are left to the jury.").

48. United States v. Vasquez, 132 S.Ct. 1532 (2012).

49. The Supreme Court uses this Latin phrase for "he said" in the doctrine governing the admissibility of expert testimony. *See, e.g.*, Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997) ("But nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert."); *see also* Hon. William G. Young & Jordan M. Singer, *Bench Presence: Toward a More Complete Model of Federal District Court Productivity*, 118 PENN ST. L. REV. 55, 72–74 (2013) (discussing due process and legitimacy).

50. *See generally* Richard H. Fallon, Jr., *Legitimacy and the Constitution*, 118 HARV. L. REV. 1787, 1794–98 (2005) (surveying legal, moral, and sociological notions of legitimacy).

51. Solomon, *supra* note 25, at 1061 (characterizing the harmless error analysis as "determin[ing] the connection between the 'error'—not a mistake but a deviation by the judge

conception requires specifying a counterfactual situation in which the trial error had not occurred.[52] If the jury's decision in the counterfactual situation is the same as in the factual situation of error, then the error would seem to be harmless. Although there are potential alternatives to this "but for" causal analysis, we suggest that it is a good place to start.

This conception may be consistent with Justice Roger Traynor's suggestion that, "an appellate court can evaluate a verdict of guilty in terms of whether there has been harmless error or harm by reference to what a rational jury might do."[53] In the oral arguments for the *Vasquez* case, Justice Alito noted that the primary reason why the Court took the case was to examine whether the harmless error test should be focused on "a rational jury or on [a] particular jury," and worried that the latter (but not the former) would force the judges to speculate.[54] In the landmark case of *Arizona v. Fulminante*, Chief Justice Rehnquist wrote for the Court that trial-level errors "may . . . be *quantitatively assessed* in the context of other evidence presented in order to determine whether its admission was harmless beyond a reasonable doubt."[55] Although provocative, Justice Rehnquist's idea has been undeveloped. Without employing a method like the one developed in this Article, it is far from clear how a reviewing court is supposed to put numbers on evidence to perform anything like a quantitative assessment.[56]

The behavioral sciences provide a starting point for thinking about this problem. For example, as of 2006, there were 48 scientific research studies, with a combined 8,474 participants, focused on the question of whether jurors are able to comply with judicial instructions to disregard inadmissible

---

or prosecution from constitutionally mandated duties to a criminal defendant—and the 'harm' (the conviction), one must look to effect's necessary antecedent, 'cause.' A breach of duty, resulting harm, and an inquiry into the causal connection . . . .").

52. David R. Dow & James Rytting, *Can Constitutional Error Be Harmless?*, 2000 UTAH. L. REV. 483, 499 (2000).

53. TRAYNOR, *supra* note 19.

54. Transcript of Oral Argument at 27, Vasquez v. United States, 635 F.3d 889 (2011) (No. 11-199), 2012 WL 950280; *see also* Chapel, *supra* note 1, at 516.

55. Arizona v. Fulminante, 499 U.S. 279, 307–08 (1991) (emphasis added). It is also important to note that in habeas contexts, the court has adopted a "substantial and injurious effect or influence" standard. Brecht v. Abrahamson, 507 U.S. 619, 623 (1993); *see also* Solomon, *supra* note 25, at 1061 (quoting Kotteakos v. United States, 328 U.S. 750, 776 (1946)) ("[I]f one cannot say, with fair assurance, after pondering all that happened without stripping the erroneous action from the whole, that the judgment was not substantially swayed by the error, it is impossible to conclude that substantial rights were not affected.").

56. *E.g.*, Jon O. Newman, *Quantifying the Standard of Proof Beyond a Reasonable Doubt: a Comment on Three Comments*, 5 LAW, PROBABILITY, & RISK 267 (2006), *available at* http://lpr.oxfordjournals.org/content/5/3-4/267.full.pdf.

evidence.[57] The optimal way to test such a question is through an experimental design, which starts with some sort of representation of a case, usually either a condensed video of a trial or a written vignette, including the essential jury instructions, and perhaps arguments from the attorneys. The researchers then create three versions of the case—one without the inadmissible evidence and two versions with it—one with an instruction to ignore the evidence and one without the instruction. Then, researchers recruit some individuals to serve as mock jurors, and they are randomly assigned to one of these three vignettes. Ideally, the subjects do not know the purpose of the experiment nor which experimental condition they are in. The researchers want to avoid giving too much attention to any one piece of evidence.

Not surprisingly, the conclusions in this sort of research vary, in part because each experiment tends to use a different case vignette. But overall, behavioral scientists can conclude that inadmissible evidence has an impact on jury verdicts.[58] Unfortunately, when litigants contest the evidence ruled admissible, it has the effect of accentuating the impact on verdicts.[59] And, judicial instructions to ignore the inadmissible evidence do not effectively eliminate its impact.[60]

One might hope that courts of appeals would consider this sort of research when considering the harmlessness of inadmissible evidence being admitted. Alas, however, this sort of evidence is rarely actually considered.[61] One problem is that the scientific literature is both too narrow in one sense and too broad in another. The literature is too narrow because it fails to conduct experiments for all the various types of errors that may occur at trial: the admission of inadmissible evidence is only one of many. The literature is too broad because even where it tests a particular type of error, the case vignette is not tailored to the particular facts and law that are presented in a particular case. Even if one can say that jurors have difficulty ignoring inadmissible

57. Nancy Steblay et al., *The Impact on Juror Verdicts of Judicial Instruction to Disregard Inadmissible Evidence: A Meta-Analysis*, 30 L. & HUM. BEHAV. 469, 475 (2006); *see also* Judy Platania & Gary Moran, *Due Process and the Death Penalty: The Role of Prosecutorial Misconduct in Closing Argument in Capital Trials*, 23 L. & HUM. BEHAV. 471, 477–78 (1999) (using videotape of penalty phase of actual capital trial to assess whether improper statements made by prosecutor impacted jury decisions).

58. Steblay et al., *supra* note 57, at 477–78.

59. *Id.* at 479–80.

60. *Id.* at 477.

61. A WestlawNext search of cases citing the Steblay article described above yielded no such instances. This study also suggests that courts "rarely rely on actual social science research about the effects of different kinds of evidence, argument, or instructions on jurors." Solomon, *supra* note 25, at 1071; *see also* Mitchell v. Gonzales, 819 P.2d 872, 877–78 (1991) (discussing reforms to jury instructions, citing jury simulation research). *But see* People v. Allen, 420 N.W.2d 499, 508–09 (Mich. 1988) (considering prior research on the topic).

evidence, one cannot say whether the erroneous admission made a difference in a particular case, which may have had much stronger (or weaker) other evidence than the typical case tested in these experiments.

To address these problems, we propose that reviewing courts should consider behavioral science experiments conducted with vignettes that represent the particular facts and law of the case presented. Although there will always be a step of inference, from these experiments on simulated jurors to the real world of properly instructed and properly informed juries, such experiments can narrow the epistemic gap.

Vignettes have become common tools in a range of scientific and practical fields including "sociology, psychology, business, and health sciences."[62] Vignette-based experiments are now published in the leading scientific journals, to predict real-world behaviors.[63] Some research has begun to explore the validity of vignette-based decisions to predict real-world behavior, and the results vary by context.[64] Jury research on such questions of validity has been indirect, but it is promising, showing that the verdicts of both mock jurors and real jurors tend to track the strength of the evidence presented.[65] A leading study in the jury research field suggests that neither "stimulus case realism" nor "study population" have a significant effect on "research conclusions"—that is, a lack of realism and a less than realistic study population do not significantly dampen applicability to real-world scenarios.[66] There are a number of theories suggesting ways in which mock juries do differ from real juries—certain variables having more or less effect in a real jury setting, difference in consequences and motivation for real and mock jurors, and different emotional responses, just to name some more

---

62. Jessica L. Collett & Ellen Childs, *Minding The Gap: Meaning, Affect, and the Potential Shortcomings of Vignettes,* 40 SOC. SCI. RES. 513, 513 (2011) (listing these fields in particular). *See generally* Rhidian Hughes, *Vignette Technique*, *in* 3 The SAGE ENCYCLOPEDIA OF SOCIAL SCIENCE RESEARCH METHODS 1184, 1184–85 (Michael S. Lewis-Beck et al. eds., 2004) (discussing the methodology and use of vignette technique).

63. *E.g.*, Aaron Kesselhim et al., *A Randomized Study of How Physicians Interpret Research Funding Disclosures*, 369 NEW ENG. J. MED. 1119, 1120–21 (2012).

64. Robert J. MacCoun, *Comparing Legal Factfinders: Real and Mock, Amateur and Professional*, 32 FLA. ST. U. L. REV. 511, 512 (2005).

65. *See* Dennis Devine, *Jury Decision Making: The State of the Science*, 7 PSYCHOL. PUB. POL'Y & L. 622 (2012) (reviewing this literature); *see also* David L. Breau & Brian Brook, *"Mock" Mock Juries: A Field Experiment On The Ecological Validity of Jury Simulations*, 31 LAW & PSYCHOL. REV. 77 (2007) (for a more direct but extremely limited test of the validity question); Neil Vidmar, *The Performance Of The American Civil Jury: An Empirical Perspective*, 40 ARIZ. L. REV. 849 (1998) (reviewing various studies that involve both surveys of real jurors and mock jury experiments, yielding similar results).

66. MacCoun, *supra* note 64, at 512 (citing Brian H. Bornstein, *The Ecological Validity of Jury Simulations: Is the Jury Still Out?*, 23 LAW & HUM. BEHAV. 75 (1999)).

salient theories; there does not appear, however, to be much empirical evidence suggesting that mock jury research is consequently inapplicable.[67] As one might expect, highly emotional situations are difficult to simulate in a written vignette. In economic contexts, some research has suggested that a lack of adequate financial incentives in simulations can lead to different outcomes than when such incentives are present.[68] For physicians' healthcare decisions, on the other hand, scholars have found a high correlation between vignette-based predictions and real-world behaviors: when a vignette predicted that the physician would do something, his likelihood of doing it was indeed five times greater.[69]

It is also important to point out that harmless error analyses are currently carried out in a form that is substantially similar to review of vignettes (except that it is polluted by the sorts of biases described in Part I above). When an appellate court makes its harmlessness determination, it commonly does so based on the record on appeal, which has been summarized in the trial court opinion, the litigants' briefs, and memoranda from court clerks. While some appellate judges may actually review the raw trial court record, even then, the judges are a step removed from the live testimony of the real trial.[70]

B. Design and Stimulus Cases

To demonstrate this method for informing harmless error analyses, we conducted a randomized controlled simulated jury experiment. The randomized controlled experiment is considered the "gold standard" for scientific evidence because it uses random assignment to control for all the possible variations that might impact an outcome ("confounds"), allowing

---

67. *See id.* at 513–17.

68. Ofer H. Azar, *Does Relative Thinking Exist in Real-World Situations? A Field Experiment with Bagels and Cream Cheese,* 49 ECON. INQUIRY 564, 566 (2011).

69. *See* Geert M.J. Rutte et al., *Measuring Physiotherapists' Guideline Adherence by Means of Clinical Vignettes: A Validation Study*, 12 J. EVALUATION CLINICAL PRAC. 491, 492 (2006) (concluding that vignettes are of acceptable validity for predicting real-world behavior); H. Sandvik, *Criterion Validity of Responses to Patient Vignettes: An Analysis Based on Management of Female Urinary Incontinence*, 27 FAM. MED. 388 (1995) (showing a correlation coefficient of 0.65, $P < 0.001$, and stating that when a vignette predicted an action, it was more than five times more likely to actually occur than when the vignette did not). *But see* D.C. Morrell & M.O. Roland, *Analysis of Referral Behaviour: Responses to Simulated Case Histories May Not Reflect Real Clinical Behavior*, 40 BRIT. J. GEN. PRAC. 182, 183 (1990) (showing that doctors' actual referral rates were not significantly correlated with their responses to vignettes, even where the case histories appeared realistic).

70. *See, e.g.*, Canter v. Koehring Co., 283 So. 2d 716 (La. 1973), *superseded by statute*, 1976 La. Acts 147 (discussing "the trial court's better capacity to evaluate live witnesses (as compared with the appellate court's access only to a cold record)").

investigators to focus on the single question at issue.[71] If a difference is observed between the experimental conditions, one can infer that the manipulation was the cause. Like tests of new drugs, we had experimental conditions representing a "control" where there was no error, and a "treatment" where an error was present.

For each case, we used a one-by-two between-subjects design, meaning that each subject saw one version of one case. We manipulated as an independent variable the presence of error (or not).[72] The primary dependent variable was whether the juror would vote to convict.

Notably, these are not surveys, which would simply ask people whether they think that the error is harmless.[73] Such an opinion poll would be subject to all the same problems of the current method of judicial speculation, and it might be even worse, since members of the public have less experience observing trials and trial outcomes. Instead, our participants are fully blinded to the purpose of the study and the existence of an error, and evaluate the randomly assigned case-version as a whole. Here, the biases can be avoided and the benchmark for comparison is the lay jury that originally decided the case. In this sense, the layperson's minimal degree of experience is a virtue, not a vice.

We selected the three criminal cases shown in Table 1. For demonstrative purposes, we handpicked one of the landmark Supreme Court cases on harmless error, *Chapman v. California*.[74] In *Chapman*, the court considered whether it was harmful for a prosecutor to draw attention to the defendant's choice not to testify.[75] The Supreme Court ruled that this type of evidence likely had an impact on the jury's verdict and was therefore harmful.[76]

To improve the generalizability of our findings, the other two cases were randomly selected from a set screened to include (a) appellate criminal cases

---

71. *See, e.g.*, TMJ Implants, Inc. v. Aetna, Inc., 498 F.3d 1175, 1195 (10th Cir. 2007) (recognizing that double-blind study is seen as the gold standard in medical industry).

72. Unlike the jury research described above, which explores the effect of admonitions to ignore inadmissible evidence, our cases were all situations where the error was not determined until the case was appealed. In such a case, if the legal question required a determination as to whether the admonition worked to secure a fair trial, there would still only be two experimental conditions: one representing the actual trial with the admonition, and one where the error was excluded altogether. In actual litigation, it is unnecessary to isolate the effect of the admonition itself.

73. Gary King et al., *Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research,* 98 AM. POL. SCI. REV. 191, 192 (2004) (explaining how survey research is subject to fundamental problems, which can be addressed through systematic use of vignettes instead, as they give a common point of reference for disparate respondents to evaluate).

74. 386 U.S. 18 (1967).

75. *Id.* at 21.

76. *Id.* at 24.

that had harmlessness as their sole issue, (b) with errors that directly involved the jury's fact-finding role, such as jury instructions or inadmissible evidence, and (c) had easily accessible and complete briefs and statements of facts. We first searched WestlawNext for "harmless error" cases decided prior to December 31, 2011, and selected the most recent 100 cases, which we then randomly sorted. We then reviewed them in order, excluding any cases that did not comport with our pre-established criteria. This process produced our other two example cases: *Lee v. Smeal*[77] and *State v. Jennings*.[78] In *Lee*, the error at issue was the improper admission of a co-defendant's confession describing the rape and double murder that were the subject of the charges.[79] This error was found to be harmless.[80] In *Jennings*, a doctor was charged with sexually assaulting three girls, and the error under analysis was whether it was proper to admit a forensic interviewer's opinion that the girls were telling the truth.[81] The reviewing court found this error harmful.[82]

From these three cases, we constructed trial vignettes from the facts of the case as presented in the appellate judge's decision and the briefs from petitioner and respondent. As shown in Table 1, these vignettes tended to be about 1,000 words long, with about 20% of that length consumed by the error. An example of our stimulus package is shown in Appendix A. Each package included the same introductory instructions, describing what their duties as jurors would entail, and providing standard instructions about the importance of considering only the evidence presented and eschewing emotion and prejudice.

The core of each condition was a two-page synopsis of the trial court record. The vignettes followed a similar pattern: each began by introducing the defendant, the charges they faced, and the fact that the defendant had entered a "not guilty" plea. After a brief description of the evidence to be presented (e.g., eyewitness testimony from two witnesses) and a summary of the prosecuting and defending attorneys' arguments, the facts gathered from court documents were presented. Finally, the participants were presented further jury instructions explaining reasonable doubt and the elements of the charges. For the ultimate charges, we used the jury instructions the defendant

77. 447 F. App'x 357, 358 (3d Cir. 2011).

78. 716 S.E.2d 91 (S.C. 2011). We planned to also utilize a third random court case and another handpicked case (five altogether), but failed to systematically manipulate the vignettes in a way that would isolate the trial court's error, which made the data unusable.

79. *Lee*, 447 F. App'x at 359.

80. *Id.* at 362.

81. *Jennings*, 716 S.E.2d at 92.

82. *Id.* at 482.

faced from that case's respective jurisdiction, using independent research as necessary where the opinions and briefs were unclear.

C. Participants, Randomization, and Instrument

During March and April 2013, we solicited jury-eligible adults via Amazon Mechanical Turk for $0.75 each, in two phases.[83] Mechanical Turk is an online population of workers willing to perform small tasks for small payments, and has become a frequently used tool for social science research.[84] All persons consented to participate according to Institutional Review Board standards.

Prior to examining any of our dependent variables, we conducted data cleaning; we excluded any uncompleted responses, as well as any response with an unrealistically short completion time, suggesting that they may not have actually read and participated in good faith.[85] We ended with 489 clean observations, spread across six conditions.

The instrument was hosted online using Qualtrics survey software. The instrument was pilot-tested with law and undergraduate students. The survey collected general demographics (sex, year of birth, race and ethnic information, level of education, household income, and zip code). After each participant read the jury instructions, the trial vignette, and the general jury instructions, the survey directed them to a specific verdict page for their assigned condition.

A demographics table is shown in Appendix B. Although more broadly representative than the populations of undergraduate psychology students that are often used for social science research, this pool does not accurately represent any particular jury venire. The study population had 330 females (48%) and 356 males (52%), aged between 18 and 80 years, with a mean of 34.8 years. The most common races reported were Caucasian (80%), African American (8%), and Asian (8%). Of our sample, 2% identified themselves as Mexican or Mexican American, and approximately 2.5% described as some other Hispanic heritage. Of our study population, 12% had a high school

---

83. In the first phase of recruitment, we randomized the respondents to the cases equally. Initial data analyses suggested insufficient statistical power for one case. So in a subsequent phase of recruitment (N=100), we assigned subjects only to *Chapman* conditions.

84. For information on the demographics of the Mechanical Turk population, see, for example, Joseph K. Goodman et al., *Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples*, 26 J. BEHAV. DECISION MAKING 213, 213 (2013); Shapiro, et al., *Using Mechanical Turk to Study Clinical Populations*, CLINICAL PSYCHOL. SCI. (Jan. 31, 2013), http://cpx.sagepub.com/content/early/2013/01/31/2167702612469015.

85. We used a 4.6 minutes cutoff, based on one standard deviation from the mean.

diploma or less education, 42% had some college credit or Associate's degrees, 35% had Bachelor's degrees, and 12% had Master's, Doctorate, or Professional degrees. Randomization successfully distributed the demographics across conditions.[86]

III. DEMONSTRATIVE RESULTS

A. Analytical Approach

For each of the three trial vignettes, there were two versions—one in which the error was present, and one in which the error was not present—which sets up a straightforward 2 (error: present or not present) by 2 (verdict: guilty or not guilty) table for analysis. If the real appellate court deemed an error to be *harmful*, then our method will cohere with that finding if the verdict rates across conditions of the trial are substantially and significantly different. On the other hand, if the error was deemed *harmless*, then the verdict rates should be statistically indistinguishable in a well-powered comparison across conditions, indicating that the presence or absence of the error had no effect. For these hypotheses tests, we provide Fisher's exact tests, coupled with computation of 95% confidence intervals for the *difference* between the two proportions.

One should distinguish a "substantial" effect from the notion of statistical "significance," because an experimenter could use a gigantic sample to provide a very precise estimate (and thus a low p-value) for a difference (say a single percentage point) that courts may not consider substantial enough to constitute harm. Our samples are not sufficiently large to raise that concern, however. Thus, we tentatively rely on the familiar statistical test that asks whether the null hypothesis (no effect at all) can be rejected, at the traditional .05 level of statistical significance.[87] We revisit these questions below.[88]

B. Sensitivity and Correspondence with Court Decisions

Because all of our cases involved multiple charges, we merged these verdicts into one variable which recorded whether our mock jurors found the

86. To confirm that demographic variations were not confounding our results, we performed multivariate logistic regressions (not shown).

87. *See, e.g.*, ROBERT M. LAWLESS ET AL., EMPIRICAL METHODS IN LAW 233 (2010) ("Traditionally, when a result has a 5 percent or less chance of occurring but occurs nonetheless, researchers consider the result to be statistically important.").

88. *See infra* Part IV.A.

defendant guilty on at least one count.[89] Referring to Figure 1, it is notable that there is a wide range of conviction rates, ranging from 22% in one condition to 68% in another condition, with four other rates in between. This variation suggests that the mock jurors were sensitive to the facts and law presented in the stimulus, and since these conviction rates are the primary inputs to our method of harmless error analysis, the method will be likewise sensitive.[90] Alternatively the respondents could have been indifferent (showing results clustered around 50/50 randomness), resolutely biased against defendants (showing results clustered around 100%) or all resolutely skeptical about the evidence (showing results clustered around 0%). On this rudimentary test of sensitivity, the method passes.

In addition to recording verdicts for each respondent, we also asked respondents to answer the question "How probable is it that [Defendant] is guilty of [the relevant charge]?" Respondents were presented with options between zero and 100%. This spectrum identified zero as "definitely not guilty" and 100 as "definitely guilty." Our jurors returned average probabilities between 39% and 73%, which further suggests that jurors were sensitive to the facts and law. Due to the nature of this variable, we did not merge these probabilities, and the statistics for each count (enumerated "Count 1," "Count 2," and "Count 3")[91] are reproduced below. Respondents were also tasked with ranking the pieces of evidence between one (the "most important piece of evidence") and five or six[92] (the "least important piece of evidence"). Because the error was not present in the control condition, an average ranking in the error condition is useful to show how jurors considered the importance of the error.

We discuss the specific hypothesis tests below, but the general trends can be observed in Figure 1. Overall, our results have some correspondence to the harmless error determinations made by real courts. In one case (*Jennings*), the court found the error not harmless, and our method concurred, showing a very big difference across conditions.[93] In another case (*Lee*), the court found the error harmless, and our method concurred, showing no detectable difference across conditions.[94] In the other case (*Chapman*), our method is not inconsistent with the finding of the court that the error was harmful, but

89. We also analyzed our data using guilty on all counts as a variable. Our hypothesis tests yielded the same results, although the base rates were different.

90. Compare this finding to the results of the single study of judges' performance. *See supra* text accompanying note 24.

91. Each case tested had three separate charges.

92. This number varies based on the case at issue.

93. State v. Jennings, 716 S.E.2d 91, 94 (S.C. 2011).

94. Lee v. Smeal, 447 F. App'x 357, 358 (3d. Cir. 2011).

leaves open the possibility that the error was in fact harmless, due to statistical uncertainty.[95]

C. Hypothesis Tests

For *Jennings*, we observed conviction rates of 22% versus 50% in the no-error and error conditions, respectively. With a substantial difference of nearly thirty percentage points, the verdict rates were significantly different ($n = 152$, $p < .001$; 95% CI:15%, 44%), and we can reject the hypothesis that the error made no difference. The 95% confidence interval provides the range of values within which the difference between the two independent population proportions lies, with 95% probability. In other words, given the observed difference of (50%-22%) = 28%, and the sample sizes used, we can be relatively confident that the true difference is anywhere from 15% to as high as 44% of the jurors being affected by the error alone, holding everything else in the case constant. While courts may disagree about how much prejudice is too much, an effect of this size would seem to be objectionable on any account. In comparing our jurors' probability rankings, we observed significant differences between probabilities for each count. Count 1 showed an 18% difference in average probability of guilt in the error condition ($n = 152$, $p < .001$; 95% CI: -30%, -10%).[96] Count 2 also showed an 18% difference in average probability ($n = 152$, $p < .001$; 95% CI: -30%, -5%). Count 3 substantially cohered with this result, with a 17% difference between averages ($n = 152$, $p < .001$; 95% CI: -30%, -5%). The jurors in the error condition ranked the erroneously admitted evidence as the second most important piece of evidence (out of five) on average and the most important piece of evidence as a median. By every measure, our results suggest that the error condition affected the respondents. Thus, our finding is in accordance with the real appellate finding that the error was not harmless.

In *Lee*, the reviewing court found the error to be harmless. Indeed, in our experiment, the proportions are nearly identical across conditions (no error = 66%; error = 68%), with confidence intervals straddling both sides of zero about equally ($n = 139$, $p = .99$; 95% CI:-14%, 17%). The top range of the confidence interval allows us to rule out the hypothesis that one out of nine jurors would have been affected by the error, but it is possible that one or two

---

95. Chapman v. California, 386 U.S. 18, 26 (1967).

96. We analyzed this variable using a Wilcoxon rank sum test. We opted to use a Wilcoxon rather than a t-test because the distributions of confidence probabilities were generally grouped into two or three distinct populations—making them parametric. It should be noted that performing a Welch's t-test on these data provided similar *p*-values.

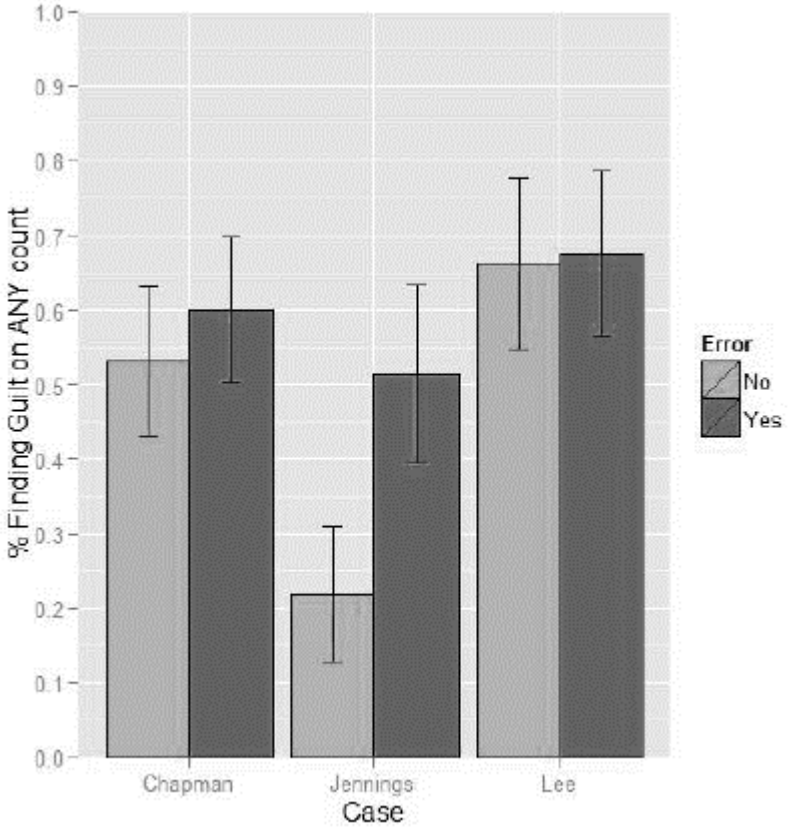
of the members of a twelve-person jury would have been affected. Count 1 showed a -3% difference in average probability of guilt in the error condition (n = 139, p = 0.63; 95% CI: -5%, 10%). Count 2 showed a 3% difference in average probability (n = 139, p = 0.34; 95% CI: -10%, 5%). Count 3 showed a 10% difference between averages (n = 139, p = 0.08; 95% CI: -2%, 5%). The jurors in the error condition ranked the erroneously admitted evidence as the third most important piece of evidence (out of six) on average and the third most important piece of evidence (out of six) as a median. Thus, while the jurors found the error to be more important than other pieces of evidence, it appears they convicted at a similar rate—and at roughly similar confidences—when it was absent. This represents an empirical demonstration of the "other overwhelming evidence" conception of the harmless error doctrine. Our best estimate is that there is no effect, and thus accords with the appellate court's finding.

The *Chapman* court found the error harmful. Our 7% observed difference between conviction rates in *Chapman* was not statistically significant at our sample size, (n = 198, p = .40; 95% CI:-7%, 21%), although the central estimate approximates nearly one out of every twelve jurors being affected. Count 1 showed a 3% difference in average probability of guilt in the error condition (n = 198, p = 0.43; 95% CI: -10%, 5%). Count 2 showed a 4% difference in average probability (n = 198, p = 0.43; 95% CI: -14%, 5%). Count 3 showed a 3% difference between averages (n = 139, p = 0.47; 95% CI: -10%, 5%). The jurors in the error condition ranked the erroneously admitted evidence as the fifth most important piece of evidence (out of six) on average and the fifth most important piece of evidence (out of six) as a median. Here, with the confidence interval ranging to a difference of 21%, we cannot rule out the possibility that the error increased the chance of a guilty verdict very substantially, making it prejudicial (not harmless) as the appellate court held. In real litigation, this case would be a good candidate for additional investment in observations.

Table 1: Case Information. This table presents information about the three cases our vignettes were based on, including the case name, the charges faced by the defendant, the relevant error addressed, and the appellate judge's decision regarding harm. Word counts are for the no-error condition / error condition.

Name	Charges	Error	Appellate Ruling	Results	Word Counts
Chapman v. California	Robbery, kidnapping, and murder under California law	Prosecutor's remarks on defendant's failure to testify	Not Harmless	Difference: 7% CI: -7%, 21%	756 / 988
State v. Jennings	Three counts of lewd behavior against children	Inclusion of testimony by a forensic interviewer	Not Harmless	Difference: 28% CI: 15%, 44%	547 / 833
Lee v. Smeal	Two counts of first-degree murder and one count of sexual assault	Inclusion of co-defendant's testimony	Harmless	Difference: 2% CI: -14%, 17%	766 / 1160

Figure 1: Mock Juror Conviction Rates by Experimental Condition. Where a judge said the error was harmful, our data would cohere with the judge's determination if there was a significant and substantial difference in conviction rates. Looking at *Jennings*, for example, the large discrepancy suggests that the error had an effect on juror verdict rates.



IV. DISCUSSION

A. Limitations

Our study had important limitations, which are also relevant to any potential use of this method for informing harmless error determinations in real cases.[97] First, we used written vignettes, which depicted the case facts and jury instructions, derived from the appellate briefs and opinion. This is a common method for social science research.[98]

In particular, a vignette-based approach may tend to over-emphasize any error. In a real trial, the erroneously-admitted piece of evidence may only consume a few minutes of a week-long trial, while it may consume 20% of a trial vignette if it is difficult to summarize the error to the same degree as other elements of the trial that may be omitted altogether. Although we have no gold-standard for comparison, it is somewhat reassuring to see in Figure 1 some cases where the error had a very large effect and some where it had little or no detectable effect. Although far from conclusive, this suggests that this method may not systematically overestimate the effects.

Nonetheless, we do not take a position with regard to what is the optimal level of investment in vignette verisimilitude. When the stakes are high, of course, greater investment will be appropriate. Real-world litigants could create more realistic stimuli in the form of longer written vignettes, videotaped trials, or even mock live trials. These changes would increase the cost of this procedure, but could improve the accuracy and perceived legitimacy of harmless error determinations. At some point, of course, so much would be invested in conducting a mock trial that the reviewing court might as well remand the case for a real retrial. Thus, if this empirical method is to have any value, it must not become a burden on the courts by duplicating the very process that it was designed to analyze.

Second, our experiment also did not allow jurors to vote in groups to render collective judgments, but instead conceives individual juror-votes as the unit of analysis. Prior research has shown that the median vote of

---

97. For an overview of the various methodological issues associated with jury simulations, see generally Shari S. Diamond, *Illuminations and Shadows from Jury Simulations*, 21 LAW & HUM. BEHAVIOR 561 (1997).

98. *See* Breau & Brook, *supra* note 65, at 79 ("Most mock 'trials' simply ask participants to read written trial transcripts or short summaries, although others have employed audio or video presentations and occasionally live trial presentations . . . . [T]he type of trial presentation medium has been found to have little effect on an experiment's outcome."); *see also* sources cited *supra* notes 57–62.

individual jurors is a good predictor of the collective jury outcome.[99] Nonetheless, deliberation is an important part of the jury process, so future studies of real trials could incorporate this element. Here again, though, demands for a gold-standard experiment start to sound like demands for a retrial, and we are unable to say where the line of "good enough" social science should be drawn in this applied setting.

Third, rather than the broad national population from which we drew mock jurors, it may be useful to recruit participants from a population that represents a particular venue in which the original case was tried. Of course, if some venues have more strict or lenient jurors, that might affect the base rates of conviction. However, the venue would not normally interact with the question of whether a particular piece of evidence or statement of the law made a difference, the question here. Our randomized method distributes demographic variations and background variables like leniency across conditions, thus eliminating any confound. Still, there may be situations where the venue is so very lenient or very strict that the error has a floor or ceiling effect, or there may be certain sorts of errors—such as pretrial publicity or a prosecutor's racially charged statement—where locality matters.

In practice, a survey firm could recruit participants from the original trial jurisdiction's jury pool. With the cooperation of the courts, litigants could even use citizens called for jury duty, testing them while they are waiting for a real trial or after they are dismissed from other cases. This is a promising route for future development.

Fourth, it may also be useful to implement more stringent participant recruitment and screening processes to simulate the process of *voir dire*. We did not exclude individuals who may have been victims of similar crimes, had strong moral objections to the death penalty, or may have had other biases. It would be relatively easy to implement a "death qualification" procedure of screening jurors, if that were thought to be important to the validity of the jury simulation.[100] In real trials, by screening out some jurors and educating the remaining ones about the case, the process of *voir dire* likely affects the judgment process of the jurors. *Voir dire* is especially important in a real trial where there will only be twelve decision-makers (or fewer in some jurisdictions for some sorts of cases). For these sorts of

---

99. *See* S. Femi Sonaike, *The Influence of Jury Deliberation on Juror Perception of Trial, Credibility, and Damage Awards*, 1978 BYU L. REV. 889, 903 (1978); Shari S. Diamond & Jonathan D. Casper, *Blindfolding the Jury to Verdict Consequences: Damages, Experts, and the Civil Jury*, 26 LAW & SOC'Y REV. 513, 545–46 (1992).

100. *See* Lockhart v. McCree, 476 U.S. 162, 173 (1986) (discussing this limitation of prior jury research).

experiments involving hundreds of jurors for a single case, the risk of a single outlier changing the outcome is minimal (and that risk is approximated by the statistical confidence intervals provided).[101] Thus further screening may not be required.

Finally, one may raise the concern that the conviction rates in the error-conditions ranged between 50% and 70%, while the real juries convicted unanimously. This may suggest that our method lacks ecological validity.[102] In a subsequent experiment, an analyst could try to improve upon the vignette by adding more emphasis to inculpatory facts. Again, the base rate of conviction is not intrinsically important to our method, unless it were so low as to create a floor-effect. Instead, we are primarily interested in the difference between the two conditions.

Overall, this method has real limitations in its reliability for estimating the causal impact of trial errors. Nonetheless, it may still be a useful supplement to the bias-laden and conclusory speculations that judges currently utilize to answer this same question. As Shari Diamond has argued more generally, "simulation, despite weaknesses, offers more reliable and valid evidence than the judge's empirically untested assumptions about jury behavior."[103]

B. The Potential Use of Experiments in Real Cases

The American system of criminal and civil litigation is based on deeply engrained adversarial norms, which make it unlikely that judges would themselves commission these sorts of harmless error experiments. Even though trial court judges have long had the power to select their own expert witnesses, the power is almost never utilized.[104] Appellate judges are likely to see the production of this sort of evidence as even more alien.

Instead, this empirical method is most likely to be used by litigants, and then presented to trial and appellate courts to support arguments for or against a finding of harmlessness. In principle, appellate courts could create new procedures for taking such evidence, perhaps utilizing special masters.

---

101. *See* Michael J. Saks, *The Smaller the Jury, The Greater the Unpredictability*, 79 JUDICATURE 263, 263–65 (1996).

102. Of course, the real jurors benefitted from a much more robust trial experience—looking the witness in the eye, seeing the photos of the crime scene, etc. Of course, these exposures may have a biasing effect that outweighs their informational effect, but that would require a profound critique of our present rules of evidence. *See* FED. R. EVID. 401–03 (defining relevant and prejudicial evidence).

103. Diamond, *supra* note 97, at 569.

104. *See* Christopher T. Robertson, *Blind Expertise*, 85 N.Y.U. L. REV. 174, 198–200 (2010) (reviewing the evidence).

However, a more familiar procedure would be to remand upon finding an error to allow the trial court to then conduct an evidentiary hearing, informed by one or more mock jury studies, for the purposes of determining harmlessness. Even now, it is not uncommon for courts to remand for the harmless error analysis.[105]

One disadvantage of remand is that the very same judge that committed the error must determine whether it was harmless. Thus, remand and reassignment might be preferable, though it would be an important and controversial change to current procedures.

Once remanded, an experiment would be conducted by an expert witness properly qualified for this sort of scientific endeavor.[106] Such an expert should rely upon the 50 years of jury simulation experiments as a foundation for that method.[107] Such an expert would need to show that his conclusions are based on sufficient facts or data and is the product of reliable principles and methods, applied to the facts of the case.[108] In other domains such as trademark law, similarly, it is not uncommon for an expert to conduct a survey or experiment specifically for the case at hand.[109]

It will be possible for such experts to bias the studies, perhaps by overemphasizing the error when designing the stimulus, for example. This problem of litigation bias is, of course, not peculiar to this particular domain of scientific evidence.[110] The pertinent question is whether the biases can be observed and managed through the adversarial process and court supervision.

Ideally, the expert would be court-appointed or blinded to avoid these biases.[111] The person who drafts the vignettes need not know, for example, whether one side, the other, or the court itself requested the study and might not even need to know what in the original trial was alleged to be the error. That element could be manipulated secondarily. Still, the randomized design of these experiments and the transparent means of conducting them, is likely

---

105. *See, e.g.*, Rosemond v. United States, No. 12–895, 2014 WL 839184, at *1 (U.S. Mar. 5, 2014).

106. *See* FED. R. EVID. 702.

107. *See* Diamond, *supra* note 97; *see also* Devine, *supra* note 65 (reviewing this literature). *See generally supra* notes 57–69.

108. *See generally* Diamond, *supra* note 97, at 563–70 (discussing the types of data that need to be collected and analyzed through the practice of a jury simulation).

109. *See generally* Reginald E. Caughey, *The Use of Public Polls, Surveys and Sampling as Evidence in Litigation, and Particularly Trademark and Unfair Competition Cases*, 44 CALIF. L. REV. 539 (1956) (discussing survey sampling to be considered as evidence in trademark litigation); Shari Seidman Diamond & David J. Franklyn, *Trademark Surveys: An Undulating Path*, 92 TEX. L. REV. 2029 (2014) (noting the importance of sampling of large numbers of people in order to record their reactions to the trademarks in suit to prove confusion).

110. *See generally* Robertson, *supra* note 104.

111. *Id.*

to reduce bias compared to the raw opinions of many other litigation experts.[112]

Any parties considering the use of such a study should, however, consider strategic aspects involved with the generation of this sort of evidence. While favorable evidence of this sort would likely be persuasive to appellate judges, the empirical (and blinded) nature of the method prevents clear predictions as to the eventual results. Litigants should make strategic considerations just as for any empirical evidence. These considerations will to some degree depend on the robustness and predictability of attorney work-product protections and, for indigent defendants, the rules around court-funding of experts.[113]

C. Burdens of Proof and Statistical Uncertainty

In a direct criminal appeal (unlike post-conviction litigation), the harmless error showing must commonly be made by the prosecutor beyond a reasonable doubt.[114] So, the relevant question on appeal in many cases is whether the prosecution can prove, beyond a reasonable doubt, that a specific error did not affect the verdict. This "beyond a reasonable doubt" standard suggests that even small effects would be worrisome, if they can be distinguished from null effects.[115]

Litigants and courts need to be quite careful about exactly what hypotheses they are able to rule out. They should be careful not to mistake a failure to detect an effect with the lack of an effect, or vice versa. And, courts should not confuse the estimated size of the effect (the marginal number of jurors

---

112. *See generally* Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66.5 J. EDUC. PSYCHOL. 688 (1974) (stating that given the option to chose between randomized and non-randomized experiments, randomized should be favored).

113. For a discussion of the relative advantages and disadvantages of introducing certain types of empirical evidence, see, for example, John H. Blume & Emily C. Paavola, *Life, Death, and Neuroimaging: The Advantages and Disadvantages of the Defense's Use of Neuroimages in Capital Cases-Lessons from the Front*, 62 MERCER L. REV. 909 (2011).

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

115. A more precise analysis would require specification of the size of the jury, and the voting rule (unanimity versus majority), both of which vary by jurisdiction and context, along with a model of jury deliberation. Still, for example, suppose the analyst has an observed difference between error and no-error conditions of 5%, which is to say for any one juror there is a 95%, likelihood that she will be unaffected by the error. The probability that *all* of the appellant's twelve jurors were unaffected is thus 95% raised to the 12th power, which is 54%, leaving a 46% chance that one would have been affected. Even much lower observed differences would raise substantial worries of prejudice. For example, even a 1% difference observed in our study would yield an 11% chance that at least one juror on the panel of twelve was affected by the error, flipping her individual pre-deliberation vote.

that are affected) from our statistical confidence that the effect is distinguishable from no effect at all (expressed by a p-value). For this reason, confidence intervals are more informative than p-values, since they show the range of possible values that cannot be ruled out at traditional levels of statistical significance.

Ideally, then, if a prosecutor commissioned such an experiment, it would yield an estimate very close to zero difference between conditions, with a confidence interval that excluded any substantial effect. If we are worried about effects of a few percent, then this will be a difficult feat for a prosecutor, requiring a very large sample size, of 10,000 or more responses.[116] This is not the first context in which it has been observed that it is nearly impossible to prove a negative.[117] This analysis could be read as expressing deep caution about ever purporting to say, beyond a reasonable doubt, that an error was harmless.

Ultimately, rather than being useful to prove harmlessness, this method may be useful for identifying the cases where an error was harmful, with a sizeable difference estimated and a confidence interval excluding zero. In the post-conviction setting, the defendant explicitly bears this burden of showing prejudice.[118] In direct appeals, this method can rebut the prosecutor's rhetorical showing of harmlessness. For this purpose, as we have demonstrated, defendants should be able to conduct reasonably-powered studies at relatively low costs.

Our study paid respondents $0.75 each, but if the courts were to make their facilities and jury pools available, this cost could be driven down to zero. To effectively power one case, a litigant might need between 150 respondents (as we used) or up to 1,000 respondents, depending on how small of an effect needs to be distinguished from no-effect. At our cost per respondent, a study would require between $115 and $750. For a more robust stimulus and allowing time for jury deliberation in person, one could pay $15 per respondent, yielding a study of 150 respondents for less than $2,500. Other costs will be required—such as the fees for a properly credentialed expert to

---

116. For example, with 5,000 subjects per condition and a 90% conviction rate in each, the 95% confidence interval for the zero difference is (-0.0119, 0.0119).

117. *See* D. A. Andow, *Negative and Positive Data, Statistical Power, and Confidence Intervals*, 2 ENVTL. BIOSAFETY RES. 75, 75 (2003) ("Negative data are data that do not enable us to reject our null hypothesis. Such data are often difficult to publish because it is not possible to prove the null hypothesis."); William C. Blackwelder, *"Proving the Null Hypothesis" in Clinical Trials*, 3 CONTROLLED CLINICAL TRIALS 345, 351–52 (1982) ("[W]e cannot actually show that therapies are equivalent, but only that the difference between them is less than a specified quantity.").

118. Blanco v. Sec'y, Fla Dep't of Corr., 688 F.3d 1211, 1238 (11th Cir. 2012).

perform the study and testify thereto. Still this method represents a relatively affordable method of evidence production.

V. CONCLUSION

Many judges, litigants, and scholars have bemoaned the harmless error doctrine, which has a profoundly important role in American jurisprudence. But aside from partisan claims that the doctrine should be applied more often, or less often, there have been no real solutions.

The availability of this new method may highlight larger questions about the harmless error doctrine. Do we really intend this doctrine to be understood causally? If not, then our willingness to affirm some convictions may express a normative commitment that it is tolerable for some errors to cause trial outcomes, notwithstanding our constitutional commitments. Notably, on the other hand, the doctrine is already distinct from that category of errors deemed "structural"—such as denial of counsel and denial of a public trial—that "affect the framework within which the trial proceeds and are not simply an error in the trial process itself."[119] These errors are not subject to harmless error analyses. One main reason for holding these errors harmful *per se* is that they are inimical at a basic level to our form of justice.

There is a plausible alternative to this causal conception of harmless error. Rather than asking whether the error affected the outcome of the trial, one could ask—and it often seems that judges are actually asking—whether the error-ridden trial reached the correct outcome. That is, is the appealing defendant actually guilty?[120] If so, then the errors can be deemed harmless, on this competing conception, even if the errors themselves were a primary cause of the conviction.[121] We emphasize that the method that we here pilot does not purport to answer this other question, and we are not confident that our method provides a reasonable method of answering that more holistic question. Importantly, we think it is the wrong question to ask, because it puts the appellate judges in the role of a substitute jury. For both criminal trials and civil trials, the Constitution vests the jury of peers with this role of

---

119. United States v. Gonzalez-Lopez, 548 U.S. 140, 149 (2006) (quoting Arizona v. Fulminante, 499 U.S. 279 (1991)).

120. *See* Simon, *supra* note 33, at 577 ("*Harrington v. California* proposes that the reviewing judge should assess whether the conviction would still have resulted in the absence of error. According to this *guilt-focused* approach, the error itself plays a marginal role.").

121. *Id.*

deciding between guilt and innocence.[122] There is no constitutional guarantee that the jury trial process will be perfect. But it should be materially so, with only immaterial errors, those that do not affect the trial outcome. It is this causal and procedural conception, rather than an outcome-oriented conception, which protects the jury's role and reflects the rule of law, in a system of distributed powers.[123] It is therefore fortunate that the proper legal question to ask is also the one that we hypothesize is amenable to empirical testing, in the way we here pilot.

Even beyond the domain of structural error, some errors that radically violate the legal protections afforded under our legal system might be deemed harmless by the experimental method we here propose. In such cases, judges could overrule our empirical determinations and call the errors harmful nonetheless. Further, Justice Harlan once recognized in dissent "that certain types of official misbehavior require reversal simply because society cannot tolerate giving final effect to a judgment tainted with such intentional misconduct.[124] In cases such as this, even if empirical evidence was presented suggesting that the error was harmless, a judge should still exercise her judgment to rule the error harmful. The specific case to which Justice Harlan referred was intentional misconduct by judges or prosecutors—and while a right to an impartial judge has been deemed a structural error,[125] judges might still rule errors harmful in their estimation absent an empirical showing when society could not tolerate such a judgment. Alternatively, the availability of this method may cause courts to develop other doctrines—perhaps a third category for reversals, alongside structural errors and harmful errors, in the causal sense we suggest. Our method need not supplant the judge's role as the ultimate decision maker. It should however inform those decisions.

One may cogently raise concerns about the fallibility of the proposed method, but it should be compared against the real-world alternative: appellate judges continuing to use something close to speculation, colored by well-known biases, to support conclusory dispositions of cases. Because the experimental method blinds respondents to the purpose of the studies and the outcome of the original trial, this method avoids those biases. And, because

---

122. The Supreme Court has cautioned that an appellate judge must not "become in effect a second jury." *See, e.g.*, Mest v. Cabot Corp., 449 F.3d 502, 516 (3d Cir. 2006) ("Generally, where reasonable minds can disagree, questions . . . are left to the jury.").

123. *See, e.g.*, Harry T. Edwards, *To Err is Human, But Not Always Harmless: When Should Legal Error Be Tolerated*, 70 N.Y.U. L. REV. 1167, 1192 (1995) (arguing that guilt-based approach to harmless error "overlooks much in its myopic fixation on factual guilt" and usurps the traditional role of the jury).

124. Chapman v. California, 386 U.S. 18, 52 n.7 (1967) (Harlan, J. dissenting).

125. Arizona v. Fulminante, 499 U.S. 279, 310 (1991) (Rehnquist, C.J.)

this method simulates both the error and no-error conditions in randomized design, it allows an inference about the cause of any difference. Even aside from whether the experimental simulation method improves the accuracy of harmless error determinations, it provides a transparent and robust method for making inferences.

APPENDIX A: EXAMPLE OF STIMULUS

The following displays the stimuli for the Chapman case, which the error manipulations shown in brackets and bold. The other vignettes are available by request.

In this case, two defendants, Ruth Chapman and Thomas Teale, are being tried for charges of robbery, kidnapping, and the murder of Billy Dean Adcock, a bartender at a local bar. In this case, you should only consider the evidence as it relates to Chapman's guilt or innocence. Both defendants have pled not guilty. Chapman claims she had no participation in either the kidnapping or the murder.

There will be testimony by two witnesses, Lawrence Niland, the owner of the Spot Club, and Mr. Montalvo, a bystander.

The prosecutor claims the defendants robbed the bar and forced Billy Adcock into their car, before finally shooting the bartender in the head and leaving him on the side of a California road.

Neither defendant has testified, and the sole defense witness was a psychiatrist brought in to assert that Chapman had episodic amnesia before the crimes started, due to a beating from Teale.

On October 18, Niland opened the bar at 9 a.m. to find the cash drawer empty and the safe unlocked. The safe still contained approximately $400, but Niland calculated there was $260 missing. Niland said the club appeared to have been ransacked, and the prosecutor claimed its condition indicated that the victim had been forced out of it.

It was established that bartender Billy Dean Adcock, who had worked at the Spot Club the night before, was seen leaving with the last customers of the night. An eyewitness, Mr. Montalvo, identified three people outside the club matching the description of the victim, Chapman, and Teale. The State presented Montalvo's testimony in order to establish that Chapman and Teale were the last customers remaining in the club the night of October 17. Montalvo testified to stopping at a light outside the club around 2:00 a.m. and seeing the victim locking the door with a woman and man near him. However, it was late at night, and the driver drove away once the light changed.

The State presented evidence that Adcock was shot in the head from close range with a .22-caliber weapon on October 18, between 2:30 a.m. and 3:00 a.m., and left in a drainage ditch in Sacramento County. It was shown that Chapman had purchased two .22 caliber guns on October 13th, five days before the murder. When Teale was arrested later, the police found a similar .22 caliber gun in his possession.

Chapman was found on October 26 in St. Joseph, Missouri. She was arrested and transported back to California, where she was held for several days. During that time she gave a detailed statement to a psychiatrist, Dr. Ralph K. Winkler, claiming she had amnesia which left her with no memory of the whole event. Blood (which was unable to be tied to a specific person) was found on Chapman's clothes, and blood matching the victim's was found on her shoes. Similar evidence connected Teale with the murder.

Teale was arrested on November 2 in New Orleans, found with a .22 caliber gun and no money. He was also driving the car which had been in the defendants' possession before the time of the murder. Blood matching the type of the victim was found on the floor mat of the vehicle in which Chapman and Teale had been traveling.

[While the defendants plead not guilty, neither defendant has testified in court. The prosecuting attorney commented extensively on petitioners' failure to explain away or challenge the evidence presented against them. Throughout the trial, the prosecutor

mentioned their silence multiple times, "Now, ladies and gentlemen… as you know, we have had no testimony at all in that regard why the guns were bought; in fact, I might add that the only person or persons that could give testimony in that regard would be, of course, the defendants themselves."]

The sole defense witness was a psychiatrist Dr. Arnold Sheuerman, who was called by Chapman in an effort to establish a defense of Chapman's amnesia, claiming she had suffered an amnesiac attack before any of the crimes, due to a traumatic beating from Teale. This witness opined that Chapman could not have formed any intention to act criminally due to a nervous condition and the admitted amnesia.

The defendants' attorney claims the State did not prove the robbery charge, as there is no evidence the missing $260 was taken by coercion or force, and that purely circumstantial evidence is not reasonable proof of misconduct by his clients. The defendants' attorney has also challenged the State's accusation of kidnapping, as an eyewitness testified to seeing the three people leaving the club, with no struggle or coercion. Finally, the defense attorney argues the gun found with Teale may be the same caliber, but not necessarily the same gun.

[The prosecutor has responded to this claim by stating, "Now . . . I don't know where that weapon is, ladies and gentlemen, and you don't know where it is, you've heard no testimony from the stand at all, and once again, the only person or persons that could tell us about where the original .22 caliber is today would be one or the other of the defendants, or both."]

The defendants' attorney claims all of the evidence is inconclusive and thus the State has failed to provide proof beyond a reasonable doubt to convict Chapman and Teale of either robbery, kidnapping, or murder.

The prosecutor concluded the State's case by claiming the evidence was enough to convict Chapman of all three charges. **[The prosecutor concluded the State's case by claiming, ". . . since you have no testimony from the stand, you must surmise from all facts and circumstances as to the exact reason why the weapon was purchased, because the only one in this room that could tell you why these guns were purchased is either one or both of the defendants. Certainly the defendant Ruth Elizabeth Chapman could tell you, she could tell you under oath, she could subject herself to cross-examination, and she could tell you then and it would be evidence before you. Once again, she has not chosen to do this."]**

APPENDIX B: DEMOGRAPHICS OF PARTICIPANTS

Our study population was slightly younger, more often male, and more educated than the U.S. Census data, available at http://www.census.gov/population/hispanic/data/2011.html. Randomization successfully distributed these variations across experimental conditions however, suggesting that race, sex, or age did not confound our experimental manipulations.

	Chapman (*n* = 198)		Jennings (*n* = 152)		Lee (*n* = 139)		Total Subjects	U.S. Census
	Error	No Error	Error	No Error	Error	No Error		
Total	100	98	70	82	71	68	489	
Panel A: Education								
Some high school	0 0.0%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	4.8%
High school graduate	15 15.0%	9 9.2%	7 10.0%	11 13.4%	4 5.6%	11 16.2%	57 11.7%	8.9%
Some college	32 32.0%	25 25.5%	25 35.7%	28 34.1%	26 36.6%	20 29.4%	156 31.9%	31.0%
Associate degree	9 9.0%	7 7.1%	8 11.4%	9 11.0%	11 15.5%	7 10.3%	51 10.4%	28.0%
Bachelor's degree	32 32.0%	40 40.8%	22 31.4%	26 31.7%	24 33.8%	22 32.4%	166 33.9%	18.0%
Master's degree	10 10.0%	13 13.3%	5 7.1%	6 7.3%	4 5.6%	7 10.3%	45 9.2%	7.0%
Professional degree	2 2.0%	2 2.0%	3 4.3%	1 1.2%	2 2.8%	1 1.5%	11 2.2%	1.4%
Doctorate degree	0 0.0%	1 1.0%	0 0.0%	1 1.2%	0 0.0%	0 0.0%	2 0.4%	1.2%
Panel B: Gender								
Male	44 44.0%	49 50.0%	38 54.3%	41 50.0%	33 46.5%	39 57.4%	244 49.9%	49.1%
Female	56 56.0%	49 50.0%	32 45.7%	41 50.0%	38 53.5%	29 42.6%	245 50.1%	50.9%
Panel C: Ages								
18-24	18 18.0%	15 15.3%	25 35.7%	21 25.6%	17 23.9%	14 20.6%	110 22.5%	9.9%
25-34	32 32.0%	45 45.9%	28 40.0%	25 30.5%	30 42.3%	26 38.2%	186 38.0%	13.3%
35-44	20 20.0%	12 12.2%	8 11.4%	15 18.3%	8 11.3%	17 25.0%	80 16.4%	13.3%
45-59	23 23.0%	20 20.4%	8 11.4%	13 15.9%	12 16.9%	8 11.8%	84 17.1%	21.0%

60+	7 7.0%	6 6.1%	1 1.4%	8 9.8%	4 5.6%	3 4.4%	29 5.9%	18.5%
Panel D: Race								
White	81 81.0%	81 83.5%	56 81.2%	63 76.8%	58 81.7%	52 76.5%	391 80.0%	72.4%
Black or African American	7 7.0%	4 4.1%	6 8.7%	7 8.5%	6 8.5%	4 5.9%	34 7.0%	12.6%
American Indian or Alaskan Native	0 0.0%	1 1.0%	0 0.0%	1 1.2%	1 1.4%	2 2.9%	5 1.0%	0.9%
Asian	9 9.0%	8 8.2%	4 5.8%	9 11.0%	5 7.0%	3 4.4%	38 7.8%	4.8%
Native Hawaiian or Other Pacific Islander	1 1.0%	2 2.0%	0 0.0%	1 1.2%	0 0.0%	2 2.9%	6 1.2%	0.2%
Some Other Race, 2+ Races, or Blank	2 2.0%	2 2.4%	5 1.7%	1 1.2%	1 1.4%	5 7.4%	15 3.1%	9.1%