THE EVOLUTION OF JURY RESEARCH METHODS: FROM HUGO MÜNSTERBERG TO THE MODERN AGE

BRIAN H. BORNSTEIN & AMY J. KLEYNHANS†

ABSTRACT

A common criticism of jury simulation research is that it usually solicits individual judgments from mock jurors in the absence of group deliberation. This Article analyzes the evolution of jury, as opposed to juror, research methods over time. Part I discusses the pros and cons of different research methods; Part II provides a historical overview of jury deliberation research; Part III analyzes recently developed techniques for conducting jury research in the computer era; and Part IV speculates about the jury research of the future. The Article’s final section offers a few concluding thoughts.

TABLE OF CONTENTS

INTRODUCTION .................................................................813
I. JURY RESEARCH METHODS: PROS AND CONS .........................817
II. JURY RESEARCH IN HISTORICAL PERSPECTIVE ..........................820
   A. Münsterberg ..............................................................820
   B. Burtt .................................................................824
   C. Weld and Danzig .........................................................826
III. JURY RESEARCH IN THE COMPUTER ERA: A NEW DAY DAWNS ...828
   A. Simulating Jury Deliberation: Findings .........................829
   B. Simulating Jury Deliberation: Pros and Cons ..................833
IV. SPECULATIONS AND PROGNOSTICATIONS ON THE FUTURE
    OF JURY RESEARCH ......................................................837
CONCLUSION ........................................................................839

INTRODUCTION

One of the most frequent criticisms of jury simulation research is that it could more accurately be described as juror research than jury research—that is, most research on the topic solicits individual judgments from mock jurors in the absence of deliberation.† It is difficult to estimate

† Brian H. Bornstein, Ph.D., is Professor of Psychology and Courtesy Professor of Law at the University of Nebraska-Lincoln; Amy J. Kleynhans, J.D., Ph.D., is a trial consultant at JurySync Litigation Consulting in Olathe, Kansas. We are grateful to Jeff Neuschatz for his feedback on the Hugo Münsterberg material covered in Part II, and to Bernard Chao, John Campbell, and Zoe Verhoeven for their efforts in organizing the 2019 Driven by Data symposium at the University of Denver Sturm College of Law.

1. The most comprehensive review of juries as groups included only 206 studies. Dennis J. Devine et al., Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups, 7
the ratio of juror research to jury research, in part because the search term “jury” identifies many studies that lack any sort of interactive group component—that is, they study juries by focusing on individual jurors. However, a nonsystematic and relatively crude measure—an inspection of recent books on jurors/juries—shows that juror research outweighs jury research by far more than a preponderance of the evidence. For example, Devine’s book on *Jury Decision Making: The State of the Science* contains one chapter on deliberation, and group processes are also discussed in a couple of other chapters out of the nine chapters total.\(^2\) Bornstein and Greene, in *The Jury Under Fire: Myth, Controversy, and Reform*, devote one of fifteen chapters, and portions of a few others, to deliberating juries.\(^3\)

Two of eleven chapters in *The Psychology of Juries*, edited by Kovera, focus primarily on juries as a deliberative body, though juries feature, to a smaller extent, in other chapters as well.\(^4\) And none of the fourteen chapters in Najdowski and Stevenson’s edited volume on *Criminal Juries in the 21st Century* deals primarily with processes and procedures involving the jury as a group, though as with the other books, deliberation appears sporadically in several chapters.\(^5\) This neglect of actual jury behavior, in the majority of what passes as jury research, persists despite frequent calls from within the research community for greater attention to the deliberation component of the process.\(^6\)

There are many reasons for this emphasis on individual juror judgments. Some are practical/statistical, and others are scientific/theoretical. From a practical perspective, mock jury studies are...
concerning the real world. A jury provides only a single data point, whereas six individual jurors provide six data points. Hence, juries have vastly reduced statistical power and require much larger sample sizes to test the effect of any jury-level experimental manipulations. This is because of the large number of variables that can affect the outcome of a jury trial. Examples of such variables include the type of case, the location of the trial, the composition of the jury, and the behavior of the jurors.

Scientifically speaking, the emphasis on individual judgments is neither a good thing nor a bad thing. Some research questions pertain more to individual reasoning processes (e.g., jurors’ comprehension of evidence). Other questions pertain more to scientific and other challenges, one prominent jury scholar has cautioned researchers: “Don’t reflexively include deliberating juries in research.”

7. For a useful discussion of the challenges involved in conducting jury research, see Norbert L. Kerr, Suggested Do’s and Don’ts for Future Jury Research: A Swan Song, in THE PSYCHOLOGY OF JURIES, supra note 4, at 257, 259.

8. Researchers could, of course, have juries deliberate with fewer than six mock jurors. Arguably, the fundamental psychological processes would be highly similar with four or five jurors as with six. However, research shows that jury size does influence the quality and quantity of deliberations. For review, see BORNESTEIN & GREENE, supra note 3, at 62–72; DIVINE, STATE OF THE SCIENCE, supra note 1, at 41–44; Michael J. Saks & Molli Weighner Marti, A Meta-Analysis of the Effects of Jury Size, 21 LAW & HUM. BEHAV. 451, 465 (1997). Nearly all of this research has compared six-person to twelve-person juries, making it impossible to know whether juries even smaller than six would function the same as or different from six-person juries. Regardless, studies employing fewer than six mock jurors would almost certainly be criticized for failing to reflect jury sizes that exist in the real world.

9. The no-show problem is particularly acute for researchers working with limited resources—that is, for basically all researchers. If one schedules six jurors and fewer than six show up, one can run the group anyway, opening the results to the criticism described in note 103, infra (while preserving the integrity of any individual juror data collected prior to deliberation); or dismiss the participants without running the study but after first compensating them, as ethical guidelines require. The only viable alternative, which is common practice, is to overschedule—that is, scheduling more than six participants, in the hope that at least six will show up. This solution, though logical, raises its own complication: what to do with any extras? They can be either be compensated and dismissed (inefficient) or included as part of an n-person jury, where n is some variable number greater than six (introduces a confound). The confound can be controlled for statistically but cannot fully account for subtle differences that might exist, say, between the group dynamics of a six-person versus an eight-person jury; it also introduces an additional variable (jury size) that would require a much larger sample size to analyze properly.

10. That is, when entire mock juries are assigned to different experimental conditions, as in comparing juries of different sizes, or in exposing juries to one of multiple versions of the same trial. See, e.g., Saks & Marti, supra note 8, at 461 (on research randomly assigning juries to different-sized groups).

11. Kerr, supra note 7, at 258. It is important to note that Kerr also raised several caveats to this conclusion, such as the facts that some research questions are only relevant at the jury level and that predictive models based on juror behavior might apply to jury behavior in some but not all cases. Id. at 265.
or instructions), whereas others pertain more to dynamic group processes (e.g., participation in discussion as a function of the group’s gender or racial/ethnic composition).\(^{12}\) Sometimes group behavior mimics individual behavior, but other times it does not.\(^{13}\) To a considerable extent, it is possible to model how a jury will behave if we have a reasonably good idea of how jurors will behave.\(^{14}\) Evidence that such modeling works comes, for example, from the robust finding that the best predictor of the jury’s verdict is the individual jurors’ predeliberation verdict preferences: “Put simply, initial majority factions usually ‘win’ deliberation.”\(^{15}\) However, modeling cannot capture everything. A great many fundamental social and cognitive psychological processes, such as social influence (e.g., normative and informational influence, persuasion), collaborative recall, the expression of individual heuristics and biases in the group context, and the precise mechanisms by which a group forges a single decision out of many individuals’ initial preferences, can only be examined by studying group deliberation. For example, deliberation might dampen or exaggerate individual jurors’ preexisting biases depending on the type of bias and other factors, such as the strength of the evidence.\(^{16}\) From a scientific perspective, then, “[w]hether it is more appropriate to study individual jurors or deliberating juries depends, more than anything else, on the underlying research question.”\(^{17}\) Looked at this way, the distinction is, in a sense, meaningless—jurors and juries are apples and oranges.

From a policy perspective, however, the distinction matters a great deal. Courts considering jury research sometimes mention the absence of deliberation and related issues having to do with a study’s “verisimilitude” as part of their justification for dismissing the research’s relevance to the


\(^{13}\) See generally \textit{Bornstein \& Greene}, supra note 3, at 11.


\(^{15}\) Devine, \textit{State of the Science}, \textit{supra} note 1, at 161. Devine qualifies this conclusion by noting that very little of the research on the relationship between individual jurors’ initial verdict preferences and juries’ ultimate verdicts has employed trials that “involve something other than one defendant facing one charge with only the two traditional verdict options available to the jury.” \textit{Id.} at 180. Other situations would include civil trials on damages, criminal trials featuring multiple charges or lesser included offenses, and nontraditional verdict options (e.g., NGRI, GBMI). In these cases, there might be no initial majority, only a plurality, which could lead to very different deliberation processes and outcomes.


\(^{17}\) \textit{Bornstein \& Greene}, \textit{supra} note 3, at 11.
issue at hand. Although other courts presented with jury simulation research have not raised the absence of deliberation or other ecological validity concerns, they often find other reasons to discount or ignore the research, if, in fact, they consider it at all.

The purpose of this Article is to summarize and analyze the evolution of jury research methods over time. Part I discusses the pros and cons of different research methods; Part II provides a historical overview of jury deliberation research; Part III analyzes recently developed techniques for conducting jury research in the computer era; and Part IV speculates about the jury research of the future. The Article’s final Part offers a few concluding thoughts.

I. JURY RESEARCH METHODS: PROS AND CONS

Any chosen research methodology has its advantages and disadvantages. Those for studying juries roughly parallel those for studying individual jurors. Bornstein identified five major methods of juror/jury research: (1) direct observation; (2) case studies/posttrial interviews; (3) archival analysis of jury verdicts; (4) experimental simulations; and (5) field studies. Each of these methods has its pros and cons. An important consideration in evaluating these pros and cons is whether the research question has to do with the jury’s outcome (i.e., verdict) or the jury’s process (i.e., individual reasoning and group interactions, collaborative decision making, etc.). Neither one is more important than the other; it depends on the research question. For example, one could ask whether allowing jurors to take notes and ask questions makes them convict more/less (an outcome question) or helps them remember the evidence and judge’s instructions better (a process question).

Direct observation requires monitoring, and typically recording, live juries in situ. The extensive record of deliberations means that direct observation lends itself well to process questions, though allowing for typically modest sample sizes, outcome questions can be addressed as

---


19. See, e.g., Rose, supra note 18, at 227, 229–30, 232 (finding that jury research from one leading outlet for such scholarship (Law and Human Behavior) was cited infrequently in appellate court opinions; when it was cited, explicit concerns about the realistic nature of the research methodology were rare, but that did not mean that the courts found the research relevant or dispositive).


21. Id. at 211.

well. Since Kalven and Zeisel published their seminal work on the American jury more than half a century ago, direct observation of juries has, with rare exceptions, been illegal. The principal exception in recent years is the Arizona Jury Project, which recorded real juries’ deliberations with the courts’ and jurors’ consent. The Arizona study provided a very rich source of data, but its scholarly output makes clear that it was labor-intensive and afforded little opportunity for experimental manipulation. Perhaps most importantly, until more judges see the value of undertaking such research, it is likely to remain an anomaly.

Posttrial interviews of jurors from one or a small number of trials are similarly labor-intensive and costly, as this technique involves locating jurors, interviewing them, and (typically) compensating them. As with direct observation, the interviews are a rich source of information, especially about process. However, the sample size is small and not necessarily representative of the population of juries, and the self-report data obtained have significant limitations (e.g., memory and self-presentation biases). These limitations are likely exacerbated when studying juries rather than jurors. That is, jurors might be reasonably good at remembering what they thought, said, and did during deliberation; but they are likely poorer at remembering what others thought, said, and did during deliberation, as well as the subtle dynamics of the group’s interactions.

Archival studies involve the creation and analysis of large databases of actual jury verdicts. As such, they are a useful source of data on jury outcomes, but they contain almost no data on jury processes, other than, possibly, relatively crude measures like the length of deliberation or procedural features like jury size or decision rule. Other process information—like what happened during deliberation—is not a matter of public record and, therefore, would not be coded for in the datasets. Moreover, the records typically lack data relevant to questions that scholars and policymakers might want to know, like the jury’s composition (e.g., race/ethnicity and gender breakdown).

Experimental simulations present mock jurors, either individually or as a group, with a hypothetical case and ask them to make decisions “as if” the case were genuine. From a scientific perspective, simulations possess a number of advantages, such as a high degree of experimental control, which confers high internal validity and permits the sorts of

25. See, e.g., Vidmar, supra note 24, at 65; Diamond et al., supra note 24, at 69.
experimental manipulations that allow for strong causal inferences. From a legal perspective, however, simulations often suffer from low ecological validity, low external validity, or both, and they almost always suffer from a lack of consequentiality. Simulations are currently the predominant method for studying both jurors and juries, and they are well-suited for studying both processes and outcomes.

Finally, field studies involve randomly assigning juries to one of multiple experimental conditions. For example, researchers have examined the effects of allowing jurors to take notes or ask questions by allowing jurors on some juries to engage in these practices and comparing them to juries that do not. In rare instances, field studies have been conducted in the context of jury observation. For example, as part of the Arizona Jury Project, some juries were instructed not to discuss the case until the deliberation phase officially commenced, whereas on other juries, jurors were allowed to discuss the case with one another prior to deliberation. Diamond et al. found that most juries allowed to discuss did so, but that being allowed to discuss the case with other jurors did not increase the rate of talking about it with non-jurors. In addition, discussion had no effect on jury verdicts, slightly reduced the amount of time they spent deliberating, and facilitated their comprehension of certain kinds of testimony.

---

26. Internal validity refers to whether the observed effects can be attributed solely to the experimental manipulations. The pros and cons of juror/jury simulation research have been debated extensively in the literature. For a recent review of the major issues therein, see Bornstein, supra note 12, at 209–14.

27. Ecological validity refers to verisimilitude—that is, how well a simulation captures the real-world situation that it supposedly simulates—whereas external validity refers to how well the simulation results would generalize to real-world behaviors. Bornstein, supra note 12, at 211–13. Ecological validity is also sometimes referred to as “mundane realism.” E.g., Kerr, supra note 7, at 266. For a deeper treatment of different aspects of validity, see Koehler & Meixner, supra note 18, at 162–63; Daniel A. Krauss & Joel D. Lieberman, Managing Different Aspects of Validity in Trial Simulation Research, in THE PSYCHOLOGY OF JURIES supra note 4, at 185.

28. The “consequentiality conundrum” refers to the fact that mock jurors’ decisions have little, if any, consequences for the hypothetical litigants or even the mock jurors themselves. Bornstein, supra note 12, at 219. So little research has directly addressed the consequentiality question that it is difficult to say whether it makes a difference in terms of jury processes or outcomes; nonetheless, it is a common reason for courts to dismiss jury simulation research. Brian H. Bornstein & Sean McCabe, Jurors of the Absurd? The Role of Consequentiality in Jury Simulation Research, 32 FLA. ST. U. L. REV. 443, 449–50, 452 (2005).

29. For review, see BORNESTEIN & GREENE, supra note 3, at 299–303; Nancy S. Marder, Answering Jurors’ Questions: Next Steps in Illinois, 41 LOY. U. CHI. L.J. 727, 737 (2010); Nicole L. Mott, The Current Debate on Juror Questions: “To Ask or Not to Ask, That Is the Question,” 78 CHI.-KENT L. REV. 1099, 1102 (2003); Penrod & Heuer, supra note 22, at 264–73.


31. Diamond et al., supra note 24, at 33.

32. Id. at 48, 62–63.
Field studies capture many of the advantages of all of the other methods, as they are high in ecological validity and often external validity as well. Unfortunately, like observational studies, they require hard-to-obtain cooperation from the courts. Random assignment becomes complicated in real-world (opposed to laboratory) settings, which, in conjunction with other uncontrollable factors, makes it hard to draw causal inferences; and some variables of interest might be impossible to manipulate in the field (e.g., whether a murder defendant is charged with a capital crime). For these reasons, field studies, while incredibly valuable, remain “difficult, costly, and rare.”

This brief overview of the advantages and disadvantages of different jury research methods illustrates why experimental simulations are the most popular approach. Although simulations have their share of drawbacks, they permit substantial experimental control and sizeable samples; they are feasible and relatively affordable; and they allow for the investigation of both process and outcome. Considering that process characteristics are one of the things that most distinguish jury studies from juror studies—that is, how individual jurors interact with one another—it is easy to see why jury researchers would gravitate toward this approach.

II. JURY RESEARCH IN HISTORICAL PERSPECTIVE

A curious irony about the overwhelming preponderance of individual jurors in contemporary psycholegal scholarship is that the earliest research on the topic focused decidedly on juries. The present Part reviews some of that pioneering research, focusing on work conducted by Hugo Münsterberg, Harold Burtt, and H.P. Weld and E.R. Danzig between 1914 and 1940.

A. Münsterberg

Hugo Münsterberg, who is generally regarded as the founder of psychology and law, was perhaps the earliest scholar to conduct jury simulation research. Although he mentioned juries only in passing in his best-known work, On the Witness Stand, a chapter of his 1914 book, Psychology and Social Sanity, was devoted to “The Mind of the Juryman.” Münsterberg begins this chapter by acknowledging the pervasive myth—then as now—that juries are irrational and capricious:

33. These factors tend to decrease an experiment’s internal validity.
34. Bornstein, supra note 12.
35. Id. at 211–13.
37. HUGO MÜNSTERBERG, ON THE WITNESS STAND: ESSAYS ON PSYCHOLOGY AND CRIME 268 (1908). For a contemporary reappraisal of Münsterberg’s pioneering work, see Bornstein & Neuschatz, supra note 36.
38. HUGO MÜNSTERBERG, PSYCHOLOGY AND SOCIAL SANITY 181 (1914) [hereinafter MÜNSTERBERG, SOCIAL SANITY]. The chapter was an expanded version of an article he published the
Every lawyer knows some good stories about some wild juries he has known, which made him shiver and doubt whether a dozen laymen ever can see a legal point. But every newspaper reader, too, remembers an abundance of cases in which the decision of the jury startled him by its absurdity.  

However, he goes on to extol the jury’s importance in administering commonsense justice: “Better to have some statutes riddled by the illogical verdicts than legal decisions severed from the sense of justice which is living in the soul of the nation.” Münsterberg touches on a number of topics that continue to be prominent in research on jury decision making: jury selection, group dynamics in jury deliberation, and jury simulation methodology. Münsterberg’s observations on jury research methodology are worth quoting at length:

If we apply these principles [of the experimental method] to the question of the jury, the task is clear. We want to find out whether the cooperation, the discussion, and the repeated voting of a number of individuals are helping or hindering them in the effort to judge correctly upon a complex situation. We must therefore artificially create a situation which brings into action the judgment, the discussion, and the vote, but if we are loyal to the idea of experimenting we must keep the experiment free from all those features of a real jury deliberation which have nothing to do with the mental action itself. Moreover, it is evident that the situations to be judged must allow a definite knowledge as to the objective truth. The experimenter must know which verdict of his voters corresponds to the real facts. Secondly, the situation must be difficult in order that a real doubt may prevail. If all the voters were on one side from the start, no discussion would be needed. Thirdly, it must be a rather complex situation in order that the judgment may be influenced by a number of motives. Only in this case will it be possible for the discussion to point out factors which the other party may have overlooked, thus giving a chance for changes of mind. All these demands must be fulfilled if the experiment is really to picture the jury function. But it would be utterly superfluous and would make the exact measurement impossible if the material on which the judgment is to be based were of the same kind of which the evidence in the courtroom is composed. The trial by jury

previous year. Hugo Münsterberg, The Mind of the Juryman with a Side-Light on Women as Jurors, 64 CENTURY ILLUSTRATED MONTHLY MAG. 711, 711 (1913).
39. Münsterberg, Social Sanity, supra note 38. On widespread myths about jury behavior generally, see Bornstein & Greene, supra note 3.
41. Münsterberg, Social Sanity, supra note 38, at 182.
42. “A jury of artisans and farmers understands and looks into a mass of neutral material as well as a jury of bankers and doctors.” Id. at 184.
43. “The mass consciousness is not an adding up, a mere summation, of the individual minds, but the creation of something entirely new.” Id. at 185.
44. Id. at 188–89.
in an actual criminal case may involve many picturesque and interesting details, but the mental act of judging is no different when the most trivial objects are chosen.\footnote{Id.} The most interesting aspect of this excerpt is its contradictory statements with respect to experimental validity. In some respects, Münsterberg argues for high verisimilitude: group discussion about a difficult, complex situation capable of engendering real doubt. Yet, in the same breath, he argues for low verisimilitude: an artificial situation, free from features of a real case that might be too distracting, and for which the right answer—“a definite knowledge as to the objective truth”—is known.\footnote{Id. at 188. Münsterberg’s preference for “objective truth” is directly contrary to the contemporary ethos in jury research, which generally holds that the standard for judging jury decisions is not whether a jury made the “correct” decision, but whether its decision was reasonable in light of the evidence. See Brian H. Bornstein & Edie Greene, Jury Decision Making: Implications for and from Psychology, 20 CURRENT DIRECTIONS IN PSYCHOL. SCI. 63, 64 (2011).} Without saying it in so many words, he seems to be arguing for external validity in the absence of much, if any, ecological validity—a tough sell in today’s empirical and policy climate.\footnote{Id. at 198.}

Ultimately, Münsterberg concludes that juries get it right most of the time and that “the social significance of the jury system ought not to be guided by the few stray cases in which the emotional response obscures the truth.”\footnote{Id. at 195–98.} Regrettably, he argues that juries’ relatively high level of performance holds only for juries composed of men.\footnote{Id. at 198.} He bases this stance on the results of the main experiment described in the chapter, one that adheres well to his description of the optimal research method for studying juries.\footnote{Id. at 189, 195.} In the experiment, he had groups of men (Harvard University students) and groups of women (advanced women students) view a series of trials in which he presented them with two cards of dark gray cardboard on which “were pasted white paper dots of different form and in an irregular order.”\footnote{Id. at 189.} Each card had between ninety-two and a hundred and eight such white dots of different sizes.\footnote{Id.} Participants’ task was to judge whether the upper card on each trial contained more, fewer, or the same number of dots as the lower card.\footnote{Id. at 189.} Participants made individual judgments (first by writing them down, then by a show of hands); discussed the matter for five minutes, attempting to arrive at unanimity; voted again; then deliberated for five more minutes and voted a third
time. Deliberation improved the votes of male groups from 52% to 78% correct, whereas female groups started and ended at 45% correct.

These results, especially the 78% versus 45% difference on the final ballot, lead Münsterberg to conclude that because “[t]he woman remains loyal to her instinctive opinion,” she is less susceptible to reason and deliberation would be less effective. Although he professes not to pass judgment—assuring the reader that “I had no prejudice in favour of or against women as members of the jury . . . I was only anxious to clear up the facts”—he can only conclude that in the interests of justice, women should not serve on juries. As he explains:

Hence we have no right to say that the one type of mind is in general better than the other. We may say only that they are different, and that this difference makes the men fit and the women unfit for the particular task which society requires from the jurymen.

Indeed, “the psychologist has every reason to be satisfied with the jury system as long as the women are kept out of it.”

As appalling as Münsterberg’s statements regarding women jurors likely seem to modern readers, it is important to place it within its historical context. Münsterberg wrote at a time when the women’s suffrage movement in the United States was in full force. Women had served on juries in several (mostly Western) states and territories at least since 1870, but those jurisdictions were the exception. The Nineteenth Amendment, giving women the right to vote, was ratified in 1920, only six years after publication of Psychology and Social Sanity, and even that did not automatically give women the right to serve on juries. They continued to be ineligible in some states and could claim an automatic exemption in others. The U.S. Supreme Court did not rule that excluding or automatically exempting women from jury service was unconstitutional until the 1970s. It seems likely that the issue of women jurors was prominent at the time, and that Münsterberg was weighing in on it.

54. Id. at 190–91. Although ten minutes of total deliberation time seems woefully short for a jury deciding a legal case, it was presumably long enough for the task with which Münsterberg’s participants were charged.
55. Id. at 194, 196–97.
56. Id. at 198.
57. Id. at 195.
58. Id. at 198.
59. Id.
60. Id. at 202.
63. Id. at 51–52.
64. Id. at 52.
Importantly, his views on gender were relatively advanced for his time in other respects. For example, he worked with women students at a time when very few of his peers did so, and he advocated for their acceptance into Harvard’s graduate program.\textsuperscript{66}

With respect to the accuracy of Münsterberg’s conclusion regarding juror gender, more recent empirical research shows that the relationship between jurors’ gender and their trial judgments and behavior is complex. Male and female jurors reach comparable verdicts in many, if not most, kinds of cases.\textsuperscript{67} However, women are relatively more likely to side with the prosecution or a civil plaintiff in cases involving child or sexual assault victims, with plaintiffs in sexual harassment claims, and with capital defendants.\textsuperscript{68} Less research has been conducted on the relationship between gender and juror behavior during deliberation, but that which has been done shows that women are less likely to be chosen as foreperson, participate less, and both see themselves and are perceived by other jurors as less persuasive.\textsuperscript{69} Although Münsterberg may have been wrong about the existence of gender differences in terms of women or men reaching “better” decisions, he was clearly right in calling attention to the subtle, yet reliable, differences in how male and female jurors make decisions—both individually and as members of deliberating groups—at trial.

\textbf{B. Burtt}

Harold Burtt published what is possibly the earliest modern law-psychology textbook in 1931.\textsuperscript{70} In his chapter on juries, entitled “Problems of the Jury and Judge,” he states that “if the case is one that goes to a jury, we must consider the extent to which this system is conducive to adequate interpretation of the evidence.”\textsuperscript{71} The sections of the chapter relevant to juries cover “the effect of discussion” and “personnel of the jury.”\textsuperscript{72} In the section on discussion, Burtt describes Münsterberg’s jury experiment and then refutes it with one of his own.\textsuperscript{73} In Burtt’s experiment, mock jurors observed a participant, who had been assigned to a “truth” or “lie” condition, being questioned about his involvement in an imaginary

\textsuperscript{66} Unfortunately, these efforts were unsuccessful. \textit{See} Bornstein & Neuschatz, supra note 36, ch. 1.

\textsuperscript{67} \textit{See generally} Tyler N. Livingston et al., \textit{Psychological Explanations of How Gender Relates to Perceptions and Outcomes at Trial}, in \textit{4 ADVANCES IN PSYCHOLOGY AND LAW} 137, 150 (Brian H. Bornstein & Monica K. Miller eds., 2019).

\textsuperscript{68} \textit{See}, e.g., Devine, \textit{STATE OF THE SCIENCE}, supra note 1, at 113.

\textsuperscript{69} \textit{Id.} at 155–56.

\textsuperscript{70} Harold E. Burtt, \textit{LEGAL PSYCHOLOGY} (1931); Thomas Grisso, \textit{The Evolution of Psychology and Law}, in \textit{THE ROOTS OF MODERN PSYCHOLOGY AND LAW} 1, 3 (Thomas Grisso & Stanley L. Brodsky eds., 2018).

\textsuperscript{71} \textit{Burtt}, supra note 70, at 157.

\textsuperscript{72} \textit{Id.} at 157, 162. The remaining sections, on judges, cover judicial prejudice associated with litigants’ race and physiognomy and judges’ attitudes, as well as their predilection for certain numbers (i.e., in sentencing criminal defendants). \textit{Id.} at 164–67.

\textsuperscript{73} \textit{See infra} notes 74–77 and accompanying text.
crime. After each interrogation, participants voted on whether the suspect was lying or telling the truth, discussed the case, and then voted again. Neither men nor women did better than chance, and discussion failed to improve their performance. Men and women were roughly equally likely to change their judgments, and when they did change, they were about as likely to change in the right direction as the wrong direction.

Burtt’s section on jury “personnel” covers a number of sources of potential bias, such as jurors’ occupation, sympathy for the underdog, and intelligence, especially when dealing with “technical and intricate matters.” All of these extralegal factors have received more recent empirical scrutiny. Thus, Münsterberg was not alone, during the first few decades of the twentieth century, in conducting jury research and in identifying topics of ongoing concern to contemporary researchers.

The field did not, however, continue on a smooth upward trajectory. After the initial flurry of activity in the early 1900s, the field went largely silent—not only the field of jury research but the field of psychology and law as a whole. Thomas Grisso, the noted historian of psychology and law, refers to the period from 1930–1960 as “the dormant era.” The reasons for this silence are too complex to detail here, but they include John Henry Wigmore’s thorough, scathing, and satirical refutation of Münsterberg’s claim that the courts should avail themselves more of psychological research; the dominance of behaviorism in American

---

74. BURTT, supra note 70, at 159–60 (citing, in part, Harold E. Burtt, Sex Differences in the Effect of Discussion, 3 J. EXPERIMENTAL PSYCHOL. 390 (1920)). A number of more recent studies have adopted this same procedure, of having lay participants judge whether a target individual is lying or telling the truth. See generally Bornstein & Neuschatz, supra note 36, chs. 5–6; Aldert Vrij et al., A Cognitive Approach to Lie Detection: A Meta-Analysis, 22 LEGAL & CRIM. PSYCHOL. 1, 5 (2015); Aldert Vrij et al., Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection, 11 PSYCHOL. SCI. PUB. INT. 89 (2010).

75. Id. at 160.
76. Id. Men were correct on 48% of their prediscussion and 47% of their postdiscussion judgments; correct judgments for women were 48% and 52%, respectively.
77. Id. Women were slightly more likely to change in the right than in the wrong direction (nine versus five), whereas men were almost evenly split (five versus six, respectively). Id. Although Burtt did not conduct statistical tests to determine whether women might have actually benefited more from deliberation than men, he does note that “[t]hese results manifestly run counter to the earlier findings of Munsterberg [sic]”—that is, Münsterberg’s conclusion that women were more likely than men to be led astray during deliberation. Id. at 161.
78. Id. at 162–64.
81. Id.
82. John H. Wigmore, Professor Muensterberg and the Psychology of Testimony, 3 ILL. L. REV. 399, 399 (1909). Wigmore’s response to Münsterberg is arguably the greatest smackdown in the history of the field. See Bornstein & Neuschatz, supra note 36, ch. 2.
psychology; courts’ continuing skepticism of social scientific research; and World War II’s demands for applying psychology in other directions.

Jury research picked up steam again in the 1960s and ’70s, at the same time that psychology and law crystalized as a coherent discipline and the American Psychology-Law Society (AP-LS) was founded. The renewed interest in the discipline, in general and in jury research in particular, was fueled by a number of forces: larger societal changes (e.g., social unrest, rising rates of crime and violence), the growth of clinical psychology as a distinct profession with forensic applications (e.g., risk assessment), and courts’ gradual acceptance of social scientific research. Around the same time, trial consulting became popular, as a result of the use of systematic jury selection in high-profile cases dealing with antiwar demonstrations and civil rights (e.g., trials of the Harrisburg Seven and Angela Davis).

The effort to provide a scientific foundation for jury selection necessarily entailed conducting research to ascertain possible relationships between jurors’ trial judgments and a variety of attitudes and demographic characteristics. Not coincidentally, the bulk of this research focuses on the relationship between individuals’ characteristics and their judgments, although there is some consideration of how those characteristics are likely to play out in deliberation (e.g., what sort of juror is more or less likely to be chosen foreman, talk a lot, or influence other jurors).

C. Weld and Danzig

A more realistic jury study than Münsterberg’s was done by Weld and Danzig, who conducted a live mock civil trial, lasting two days, before a twelve-person, all-male jury, a twelve-person all-female jury, and a seventeen-person mixed-gender jury, as part of a moot court exercise.

84. Grisso, supra note 70, at 6.
86. See generally Bornstein & Neuschatz, supra note 36; Grisso, supra note 70, at 8–9.
87. See Hans & Vidmar, supra note 62, at 81; Reid Hastie et al., Inside the Jury 124 (1983).
manner of a real trial, with several minor exceptions.” The case involved a hotel’s (the defendant) potential liability for the theft of a patron’s (the plaintiff) cash and jewelry by a nonemployee who posed as a front desk clerk and took the valuables for safekeeping. Each party presented three witnesses, and the case concluded with closing arguments and judge’s instructions.

Mock jurors, who were Cornell University psychology students, rated their own verdict leaning on a nine-point scale, at eighteen different stages of the trial (e.g., after direct and cross-examination of each witness). Each jury then retired to a separate room, took an initial yes/no vote on liability, and deliberated for up to thirty minutes, taking additional votes as necessary until reaching unanimity. This method of having mock jurors update their verdict preferences throughout the trial is somewhat rare in contemporary jury research—most studies simply ask for a single verdict after all evidence and jury instructions have been presented—but it is not unheard of. The technique has its advantages and disadvantages. On the positive side, it almost certainly corresponds to the ebb and flow of jurors’ actual reasoning processes and permits a finer grained analysis of their response to individual pieces of evidence or argument. On the negative side, it is contrary to actual practice—jurors might naturally shift their leanings throughout trial, but they do not make explicit judgments until they receive their charge from the judge—and it raises the risk that initial, explicit judgments will bias their subsequent ones.

Three of Weld and Danzig’s findings are particularly noteworthy. First, although “the jurors did not reach decisions through a logical analysis of the case … the juries as a whole were well aware of the important issues in the case,” and their judgments throughout the trial tracked the orientation of the evidence. Second, there were no gender differences, which stands in stark contrast to the much less realistic

89. Weld & Danzig, supra note 88. The principal exceptions were the absence of voir dire and swearing in the jurors. They reported that “[t]he excellent acting by the witnesses, the introduction in evidence of photographs and blueprints, and the earnestness of all participants in the trial gave an air of reality to the proceedings such as to simulate an actual suit at law.” Id.
80. See id. at 519–20.
81. Id. at 520–28.
82. See id. at 518–20.
83. The scale, which combined the two verdict options with varying degrees of confidence, ranged from “conviction that the defendant is not liable for damages” to “conviction that the defendant is liable.” Id. at 519.
84. See generally Bornstein & Greene, supra note 3, at 12.
85. Evidence suggests that people have a tendency to stay committed to an earlier decision once they have made one. See Jonathan Baron, Thinking and Deciding 192–226 (4th ed. 2008). This tendency could explain, in part, why the vast majority of jury verdicts reflect the majority of jurors’ predeliberation preferences—that is, jurors tend to stick with the preliminary decisions with which they enter deliberation.
86. Weld & Danzig, supra note 88, at 534.
deliberation study conducted by Münsterberg. None of the juries reached a unanimous verdict, but all three favored the defendant at the end of deliberation: 10–2 for the all-male jury, 8–4 for the all-female jury, and 11–6 for the mixed jury. Combining all three juries, thirteen of nineteen women (68.4%) ultimately voted for the defendant, as did sixteen of twenty-two men (72.7%). And third, deliberation did little to change jurors’ predeliberation leanings. Indeed, only four jurors’ initial verdicts in the jury room diverged from their final judgments during trial, and only one juror (out of forty-one total) changed his vote in the course of deliberation (i.e., from one straw poll to a subsequent one).

III. JURY RESEARCH IN THE COMPUTER ERA: A NEW DAY DAWNS

Like much psychological research, many jury simulation experiments are currently administered via computer rather than using the old-fashioned, paper-and-pencil method. Results of these methods are largely comparable, including for studies of individual juror decision making. Computerized administration is fairly straightforward in the lab, where experimenters can maintain a high degree of experimental control. The situation becomes a good deal trickier when participants complete studies online from a remote location. In these situations, experimenters can monitor participants’ behavior to some extent, using relatively crude measures such as reaction time and questions gauging whether participants are really paying attention, but they have no way to detect whether participants are, say, drinking alcohol, watching television, talking on the phone, or seeking input from others while completing the study. Computer administration also makes it easier than it was for previous generations of psychological researchers to reach nonstudent samples, which increases the representativeness of the sample—a significant concern for mock jury research.

97. See supra text accompanying notes 48–60.
98. Weld & Danzig, supra note 88, at 532.
99. See id. at 534. Weld and Danzig refer to their participants as “girls” and “boys”; we have changed the terminology to “women” and “men” to reflect more modern sensibilities.
100. See id. at 532. The tendency for jury (group) verdicts to reflect the majority of individual (juror) judgments is a robust finding in the literature. See DEVINE, STATE OF THE SCIENCE, supra note 1, at 3.
102. See, e.g., Kevin O’Neil et al., Web-Based Research: Methodological Variables’ Effects on Dropout and Results, 35 BEHAV. RES. METHODS INSTRUMENTS & COMPUTERS 217, 217 (2003).
103. The most widely used platform for recruiting nonstudent participants is Amazon’s Mechanical Turk (MTurk). Considerable research has investigated the validity of MTurk data. Although early findings suggested that MTurk and other data were largely comparable (see, e.g., FRANK R. BENTLEY ET AL., COMPARING THE RELIABILITY OF AMAZON MECHANICAL TURK AND SURVEY MONKEY TO TRADITIONAL MARKET RESEARCH SURVEYS (2017), http://web.mit.edu/bentley/www/papers/chi-case-surveys.pdf), more recent work has raised serious concerns, such as the possibility that much MTurk data is generated by bots. Jesse Chandler et. al., Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers, 46 BEHAV. RES. METHODS 112, 112–13 (2013). A full discussion of the pros and cons of MTurk and similar online platforms is beyond the scope of the present paper. On the use of student...
Online platforms could, of course, allow mock jurors in multiple locations to watch a trial and then deliberate together in real time, similar to the way that multiple players can participate in role-playing games, card games, or fantasy football leagues. Such a system would expand the diversity of the jury pool, allow “players” to sign in on their own schedule and participate from their preferred location, and involve face-to-face interaction. It also seems feasible with existing technology. However, it still requires multiple live participants to engage in the task at the same time, which could create scheduling obstacles.

A. Simulating Jury Deliberation: Findings

These obstacles can be overcome by simulating the group component—that is, by leading individual mock jurors to believe that they are engaging in a group discussion, when, in reality, the other jurors are a computer program that provides prefabricated responses. For example, Salerno and Peter-Hagene developed a procedure in which individual mock jurors (jury-eligible undergraduates who received course credit) took part in a computer-mediated mock jury experiment.104 Participants were informed that they would be randomly assigned to six-person juries composed of people participating concurrently in two different computer laboratories on campus.105 Each participant read a summary of a murder trial, with accompanying photographs and jury instructions, on individual computer monitors.106 They were ostensibly then “assigned” to a jury with five other participants, with all of the mock jurors participating in an online deliberation by typing in their comments; in reality, the comments of the other five “jurors” adhered to a preprogrammed script.107

Participants reported their verdict preference, and confidence in their verdict, at three time points: predeliberation; after the first round of deliberation comments, which included no experimental manipulations; and at the end of seven additional rounds of deliberation (postdeliberation).108 Four of the confederate jurors consistently agreed with the participant, and one (the “holdout” juror) consistently

105. Id. at 584.
106. Id. at 583. The trial was based on an Australian case in which a man was tried for murdering his wife by slitting her throat. R v Velevski [2000] NSWCCA 445 (Austl.).
108. Id. at 584–86.
disagreed. The experimental manipulations were implemented in the holdout juror’s comments from the second round onward. Specifically, the juror was portrayed as either a man (Jason) or a woman (Alicia), whose comments reflected one of three different emotions: anger, fear, or none.

Salerno and Peter-Hagene included several features to enhance the procedure’s realism and make participants believe that they were genuinely interacting with “live” participants. They checked participants’ identification numbers to ensure they were in the correct room; staged incoming phone calls with the other lab; showed participants an animated progress wheel while the juries were being formed, followed by a display with six usernames, including their own; drew the confederates’ comments from a pilot study, in which real college students responded to the case in their own words; and had confederates refer to one another and the participant directly by username. These strategies were apparently successful, as only 20 of 244 participants (8%) reported suspicion that the interaction was fake. The authors obtained roughly comparable suspicion rates in two subsequent experiments: 14% using a diverse adult sample drawn from Amazon’s Mechanical Turk (MTurk), and 5.5% using an undergraduate student sample. In light of the relatively low number of participants who (correctly) suspect that their fellow jurors are not real people, the computer-mediated mock deliberation paradigm appears to be an effective way of simulating jury deliberation.

Kleynhans recently adapted this simulated jury deliberation procedure to examine whether jurors would report the deliberation misconduct of a fellow juror. The research was inspired by recent cases addressing whether, after a verdict has been rendered, jurors may testify about deliberation room racism that may have affected the validity of the

---

109. To increase realism and the likelihood that the participant would change his or her verdict, one of the majority confederates changed verdicts to side with the holdout after four rounds. Id. at 585.
110. Id. at 583.
111. E.g., “Seriously, this just makes me angry . . .,” “Ok, this is getting really frustrating . . .,” “ugh this whole thing really creeps me out . . .,” “it scares me to think about how . . . .” Id. The study’s major substantive finding was that participants became less confident in their original verdict after a male holdout expressed anger, but more confident in their original verdict after a female holdout expressed anger; holdouts exerted no effect on participants’ judgments when they expressed no emotion or fear. Id. at 586–87. For a similar example of how a preprogrammed minority group member’s position can lead others (actual participants) to change their opinion in a computer-mediated discussion, but as part of a nonjury task, see Félix Moral-Toranzo et al., Anonymity Effects in Computer-Mediated Communication in the Case of Minority Influence, 23 COMPUTERS HUM. BEHAV. 1660, 1660 (2007).
112. Salerno & Peter-Hagene, supra note 10, at 583–85.
113. Id. at 583. These participants were excluded from data analyses, along with an additional fourteen participants (6%) who failed the holdout gender manipulation check.
verdict.\textsuperscript{116} Prior to the 2017 Supreme Court decision in \textit{Peña-Rodriguez v. Colorado},\textsuperscript{117} such testimony was only admissible in many jurisdictions if the conduct was reported before the verdict.\textsuperscript{118} Therefore, courts were relying on jurors both to report any racism and to do so before the conclusion of their jury service.\textsuperscript{119} With no admissible evidence of racial misconduct, defendants would be unable to meet their burden to receive a new trial.\textsuperscript{120} In \textit{Peña-Rodriguez}, the Court held that the Sixth Amendment right to a fair and impartial jury required the admissibility of juror racial bias testimony regardless of when the reporting occurred.\textsuperscript{121} While timing is now no longer relevant to admissibility, jurors still need to report the misconduct in order for a defendant to have recourse. The following study was designed to examine whether, and under what conditions, jurors would report the racist misconduct of a fellow juror.

The first study took place in a controlled lab environment, and the second was conducted online through MTurk.\textsuperscript{122} In the lab study, university students signed up for a study that they were told would be examining different deliberation procedures, either online or in person. The actual study involved only online participation, but the fabricated purpose of the study was designed to mitigate participant suspicion as to why they would be deliberating through a computer instead of face-to-face. There were four available participant time slots for each session, and when participants arrived at the lab they saw other students waiting to participate. This was intended to strengthen participants’ belief that they would be deliberating with others in real time. Participants were taken into separate, private computer rooms; were prompted to enter their basic demographic information; and then presented with the facts and evidence of a murder trial.

At the conclusion of the evidence, participants entered their initial verdict and were asked to provide a rationale for their decision to be viewed by the other participants. To enhance the realism, the computer then informed participants that the program was waiting for all responses to be recorded by the other participants. Following a short wait, researcher-prepared responses were revealed, alongside the demographics of each juror. The verdicts and rationales were a mix of guilty and not guilty, but one juror’s rationale included racially biased remarks about the black defendant. Participants were asked to provide a response to the other


\textsuperscript{117} 137 S. Ct. at 858.

\textsuperscript{118} Id. at 858–59.


\textsuperscript{120} Id.

\textsuperscript{121} Peña-Rodriguez, 137 S. Ct. at 869.

\textsuperscript{122} See Kleynhans, supra note 115, at 60–61.
jurors’ comments and then viewed another set of comments from their fellow jurors; they then provided their final verdicts. Throughout the study, participants had the ability to ask any questions or inform the research assistant of any concerns either through their computer on Skype or directly to the assistant, who was located in one of the adjoining rooms.

Two hundred and three students participated in the lab study; only seven (3%) were removed for suspecting that their fellow jurors were computer-simulated.\footnote{Id. at 45, 50.} Four of these suspicious students told the research assistant that they did not believe that another “real” student at their school would say something so blatantly racist. The other three noted their suspicion that the other participants were not real, but they did not specify why.

In the second study, Kleynhans and Bornstein attempted to replicate their results outside of a university context with a purely online community sample. Participants signed up through MTurk for a study that was advertised as intending to examine the ways in which jurors make verdict decisions. They were told that the study was being run live, with a researcher available via email for any questions or concerns. On each page throughout the study, a text box offered participants the ability to communicate any thoughts or concerns with the researcher. All procedures were otherwise the same as in the first study. One hundred and ninety-four participants completed the study, fourteen of whom were removed for suspicion (7%).\footnote{Id. at 62.} Three of the fourteen noted that their suspicion was specifically because of the shocking nature of the racist remark.

The higher incidence of suspicious jurors in the second study is likely due to its solely online nature. In the first study, participants were required to visit the lab, saw other potential participants arrive for the same start time, and were aware that these other participants were sitting in adjacent rooms. These factors enhance the realism and lower the likelihood that participants will question whether they are deliberating with real other jurors. It could also be that as “repeat players,” MTurk workers are savvier than college undergraduates by virtue of their greater familiarity with the methods and measures being used.\footnote{Chandler et. al., supra note 103, at 114.} Indeed, this is one of the main criticisms against MTurk and other crowd-sourced samples.\footnote{See id.} Although some research has found that, compared to more traditional convenience samples, MTurk workers are somewhat more diverse and no more likely to engage in problematic responding,\footnote{E.g., Elizabeth M. Briones & Grant Benham, An Examination of the Equivalency of Self-Report Measures Obtained from Crowdsourced Versus Undergraduate Student Samples, 49 BEHAV. RES. METHODS 320, 332 (2016) (showing data obtained from MTurk workers and a second online recruitment platform were statistically equivalent to data obtained from an undergraduate student}
concerns about the integrity of MTurk data. The more problematic of those concerns are that MTurk workers deliberately provide fraudulent data and that a nontrivial proportion of the data comes not from human workers at all but from malware or “bots” that respond randomly.  

B. Simulating Jury Deliberation: Pros and Cons

The results of the handful of computer-mediated deliberation studies suggest several conclusions. First, with suspicion rates across five studies ranging from 3%—14%, participants in both studies overwhelmingly accepted that they were deliberating in real time with other participants. The low suspicion rate illustrates that simulated jury deliberations are a promising means of studying deliberation behavior while reducing the practical and logistical complications involved in gathering six or more mock jurors together in the same place at the same time.

Second, the method’s plausibility appears to vary depending on sample, procedural, and stimulus characteristics. Salerno and colleagues,129 and Kleynhans and Bornstein,130 both found higher suspicion rates for MTurk than for college student samples. Student mock jurors are necessarily more homogeneous than nonstudents, especially in terms of age, occupation, intelligence/education level, and potentially in other respects as well.131 Despite regular concerns in the jury research community about the use of student mock jurors, empirical findings suggest that differences between student and nonstudent mock jurors typically range from nonexistent to small and inconsistent.132 Moreover, there is evidence that the use of students in jury research is declining over time,133 possibly due to the advent of MTurk and similar research sample); Michael Buhrmester et al., *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?*, 6 PERSP. ON PSYCHOL. SCI. 3, 4 (2011) (showing MTurk samples were more demographically diverse than standard Internet and college samples and provided data as reliable as those obtained via traditional methods); Elizabeth A. Necka et al., *Measuring the Prevalence of Problematic Respondent Behaviors Among MTurk, Campus, and Community Participants*, PLOS ONE, Jun. 2016, at 1, 13 (showing MTurk workers engaged in problematic responding, based on self-report, at roughly the same rates as student and community samples).


129. See supra note 104 and accompanying text.

130. See Kleynhans & Bornstein, supra note 119, at 196–97.

131. E.g., students and nonstudents might differ, on average, in terms of race/ethnicity and socioeconomic status, though undergraduate samples vary widely in these respects across geographical regions (e.g., upper Midwestern versus Southwestern United States) and even within regions depending on the particular institution (e.g., a highly selective private college versus a less selective public university). They might also differ in terms of their personality and attitudes. See generally Bornstein et al., *Sampling Issues*, supra note 103, at 14.


platforms that make it easier to recruit—and less expensive to compensate—nonstudent samples.

Procedural variations are more likely candidates for the observed student-nonstudent difference than the samples’ demographic characteristics. Although both research teams strove to hold the procedure constant in studies using student and MTurk samples, confounds necessarily arose. Specifically, students participated in a research laboratory on campus, whereas MTurk workers participated in an unknown location; students interacted with a live experimenter, whereas MTurk workers either had no interaction with the experimenter or had the option of communicating with the experimenter electronically;\textsuperscript{134} students were monitored (at least loosely) while participating, whereas MTurk workers were not; and students believed their fellow jurors were like them (i.e., students in another lab on campus), whereas MTurk workers’ fellow jurors (had they really existed) could have been anyone, anywhere. These confounds are all fairly subtle, but when taken together and in conjunction with MTurk workers’ high degree of familiarity with online studies, they could explain why the paradigm seemed less plausible to them than to student participants.

It is also worth noting that a subset of participants in the Kleynhans and Bornstein studies, regardless of whether they were students or MTurk workers, became suspicious of the computer-mediated deliberation precisely because of the experimental manipulation—that is, the blatantly racist comments of one of the confederates. Moreover, the racist comment was more likely to cause the simulation to fail when it asked students to believe that a fellow student made the comment (four of seven mentioned it) than when the simulation asked MTurk workers to believe that a fellow worker made the comment (three of fourteen mentioned it).\textsuperscript{135} These findings lead to two conclusions: first, the plausibility of computer-mediated deliberation depends, to some extent, on the content of the “canned” confederate comments; and second, the plausibility of certain comments might vary depending on other procedural or sample characteristics.\textsuperscript{136}

Third, although the low suspicion rate supports the utility of the method, insofar as suspicious participants need to be excluded from analysis, it is not without costs. At the upper end, a suspicion rate of 14%\textsuperscript{137}—which could be even higher for certain combinations of

\textsuperscript{134} Salerno et al.’s MTurk study apparently involved no communication in real time between the participant and the experimenter; Kleynhans and Bornstein’s MTurk study included an electronic communication option, though relatively few participants took advantage of it. See Salerno et al., supra note 114, at 65.

\textsuperscript{135} See Kleynhans, supra note 115, at 50.

\textsuperscript{136} To disambiguate whether the procedural differences or the sample differences are driving such effects, future research would have to bring a more diverse, MTurk-like sample into the lab or have a widely dispersed student sample engage in an MTurk-like procedure.

\textsuperscript{137} Based on data from Salerno et al., supra note 114, at 63.
participants, procedures, and trials—means that roughly one of every seven participants must be replaced. The resultant costs of doing so, in terms of time and participant compensation, are not trivial. Perhaps more problematic, there are implications of excluding this subset of participants from analysis. Excluded and retained participants clearly differed in that the excluded participants were smart enough, attentive enough, or simply skeptical enough to realize the procedure was a sham. These traits could be correlated with any number of behaviors relevant to the excluded participants’ conduct during deliberation, such as their receptivity to counterarguments, memory for evidence, and willingness to take a minority or leadership position (e.g., foreperson), as well as the quality and quantity of their contributions to the discussion. Thus, removing suspicious mock jurors could potentially limit the representativeness of the sample.

A potentially intractable problem with computer-mediated deliberation is that mock jurors are interacting remotely rather than face-to-face. It is a well-documented psychological finding that the degree of temporal, spatial, or social distance between persons predicts a host of cognitive, social, and behavioral differences. Of perhaps greatest relevance to the jury context, greater spatial distance (perceived or actual) is associated with higher level, more abstract representations of events; these abstract representations entail less detail, less contextualization, and a greater focus on abstract concepts like morality. At a more social level, greater distance would almost necessarily interfere with bonding among the mock jurors, which real jurors describe as central to their task and generally a positive experience. The relative anonymity of computer-mediated communication should diminish group members’ sense of group identity, which could free them to make more extreme comments, while also making it less likely that one group member would influence any other group members.

Some research has addressed the topic of online versus face-to-face group decision making, albeit not in the jury context. For example, Thatcher and de la Cour assigned four-to-seven-person groups a series of

---


140. Disruptions to the juror bonding process, as when jurors have heated disagreements, can sometimes be a source of stress. See generally BORNSTEIN & GREENE, supra note 3, at 17–36; Monica K. Miller & Brian H. Bornstein, The Experience of Jurors: Reducing Stress and Enhancing Satisfaction, in STRESS, TRAUMA, AND WELLBEING IN THE LEGAL SYSTEM 247, 247 (Monica K. Miller & Brian H. Bornstein eds., 2013).

141. See generally Moral-Toranzo et al., supra note 111. For evidence that anonymity both liberates one to take an extreme position and decreases the likelihood that others will find the extreme position persuasive, one need look no further than the Internet.
decision tasks; half the tasks were done face-to-face and half via a computer-mediated support system. The decisions were work-related problems like those that might be used for entry-level employees during a training exercise. The decision format (face-to-face versus computer-mediated) did not affect group members’ performance on a number of group process variables: leadership, initiative, judgment, or interpersonal sensitivity. It is very easy to see how these variables would be relevant to the process of jury deliberation (e.g., the task demands judgment, and effective deliberation requires individual jurors to demonstrate initiative). The researchers also measured participants’ satisfaction with both the group’s decision outcomes and the decision-making process, especially in terms of format. Satisfaction with the decision outcome and the decision process were both consistently higher for the face-to-face than the computer-mediated condition. If jurors—even mock jurors—are less satisfied with computer-mediated than with face-to-face deliberation, it could have significant and far-reaching implications in several respects. First, jury evasion is a very real and costly problem; more satisfied jurors can help counteract the myth that people should avoid jury duty at all costs. Second, serving as a juror has broader implications for civic attitudes and engagement, such as a greater tendency to vote. Third, juror dissatisfaction could play out in any number of subtle ways during deliberation, which could compromise the validity of computer-mediated deliberation studies.

Overall, computer-mediated deliberation, like most innovations, solves some existing problems while creating new ones. On the one hand, it eases some of the practical and methodological restrictions of standard jury research, and it demonstrably works for the majority of participants. By simulating jury deliberation, it allows researchers to collect data from individuals as if they were members of a group. The paradigm generates

---

143. Id. at 210–11. Leadership measured emergent leadership during the discussion; initiative measured propensity to participate in the discussion; judgment measured ability to evaluate data and reach logical decisions; and interpersonal sensitivity measured awareness and consideration of others in the group. Id. at 208. Moreover, the relationships between these group process variables and participants’ personality traits—the main variables of interest—were the same regardless of decision format. Id. at 210–11.
144. Outcome: “How satisfied were you with the decisions agreed upon?” Process: “How satisfied were you with the process used (either [face-to-face] or [computer-mediated])?” Id. at 208.
145. Id. at 211–12.
146. See BORNSTEIN & GREENE, supra note 3, at 17–24.
147. Id. at 9, 29–31. The relationship between jury service and voting is strongest for criminal jurors and civil jurors who serve on twelve-person juries or juries with a unanimous decision rule. See JOHN GASTIL ET AL., THE JURY AND DEMOCRACY: HOW JURY DELIBERATION PROMOTES CIVIC ENGAGEMENT AND POLITICAL PARTICIPATION (2010); Valerie P. Hans et al., Deliberative Democracy and the American Civil Jury, 11 J. EMPIRICAL LEGAL STUD. 697, 697–99 (2014).
148. As described in Part IV, infra, more advanced technologies might be able to offset some of these drawbacks.
rich data that would be impossible to obtain in traditional (individual) juror research, while avoiding the costs and complications inherent in traditional (group) jury research. On the other hand, there are signs that it may work better in some situations than others. Certain participant samples, means of administration, and kinds of trials or juror comments appear more likely to arouse suspicion than others. Excluding suspicious participants, even when they are in a relatively small minority, raises concerns about sample representativeness. Of course, computer-mediated deliberation is a fairly new methodology; additional research will no doubt continue to refine the procedures and answer questions about the kinds of research questions for which it is best suited.

IV. SPECULATIONS AND PROGNOSTICATIONS ON THE FUTURE OF JURY RESEARCH

Technological advances have always helped drive scientific theory as well as methods, whether the particular technology is the printing press, the microscope, the telescope, the computer, or the particle accelerator. What does current cutting-edge technology bode for the field of jury research? Looking ahead to the future necessarily entails a degree of speculation; with that caveat in mind, we consider a couple of modifications to the computer-mediated deliberation paradigm as well as one more radical departure from the current status quo.

One modification to the computer-mediated deliberation paradigm, as described above, would be to incorporate it with artificially intelligent software. Although researchers using what we refer to as the “one-participant” paradigm took pains to make confederates’ comments sound natural and like they came from actual participants’ peers, the comments were still constrained by the fact that they were identical regardless of what participants said themselves. This constancy is desirable in terms of experimental control, but it could introduce artificiality if, depending on what participants said, it sounded like nonnatural discourse. Perhaps more importantly, the procedure does not allow for a multi-stage dialogue, wherein the participant enters a response, the computerized confederate comments on the participant’s response, the participant then refers to the confederate’s comment in her own rejoinder, the confederate replies to the rejoinder, and so forth. For example, in the Salerno and Peter-Hagene study, after the holdout juror expressed anger initially, his or her next comment could vary depending on whether or not the participant referenced the emotion (e.g., “I don’t understand why it makes you so angry”). A more sophisticated program, drawing on artificial intelligence,

---

149. See supra note 103 and accompanying text.
150. See supra Section III.A.
151. As described in Part I, supra, this trade-off between experimental control (or internal validity) and ecological validity is a recurrent challenge in juror/jury simulation research.
152. See Salerno & Peter-Hagene, supra note 104, at 586.
could produce more nuanced responses by the confederates, which would provide greater realism (thereby reducing suspicion) while also generating a richer dataset.  

A second modification to the computer-mediated deliberation paradigm would involve online deliberation among six (or more) genuine participants instead of among one genuine participant and five preprogrammed confederates. The deliberation would be akin to a private chat room with only the mock jurors as the participants. This more expansive computerized paradigm would have several advantages: real participants could respond more directly and flexibly to one another than a single participant could respond to “canned” comments; their comments would necessarily sound more natural than programmed responses, thereby reducing the suspicion problem; and a live confederate could still introduce any desired experimental manipulations, while having the flexibility to adapt (within set constraints) to the evolving contours of the discussion. There would, of course, be disadvantages also: the need for additional computers and laboratory space if run on campus; the compounding of MTurk (or similar remote, online platforms) problems if run off campus; scheduling challenges regardless of where the research took place; and diminished (mock) juror satisfaction when participating via computer rather than face-to-face. Video hookups might be able to alleviate this last concern, but they also present greater hardware and software demands. In addition, beyond a modest number of jurors (e.g., six or so), it would likely be difficult to fit all the deliberating jury members on a single screen.

Another, somewhat more extreme means of enhancing ecological validity would be to place mock jurors in a fully simulated jury room via virtual reality (VR). VR technology is perhaps best known for its use in immersive game playing, but it has myriad other uses as well, being employed extensively in educational/training, clinical/therapeutic, and research contexts. Some of that research investigates characteristics of the VR experience itself, whereas other research uses VR as a tool for

153. We do not intend this as a criticism of the Salerno & Peter-Hagene study. Although we assume that the more sophisticated software of the sort we describe is already available—or soon will be—we could by no means design it ourselves.

154. In contrast to the “one-participant” variant, described supra, we might call this the “multiple-participant” variant.

155. Having participated in video conferences with participants in multiple locations, we can attest that once the number gets past four or five, it becomes difficult to see everyone at once and keep track of who is saying what.


investigating other substantive topics across a wide variety of scientific disciplines. One recent paper identified 21,667 articles about VR in the Web of Science database; the most frequent subject categories were computer science, engineering, psychology, neurosciences and neurology, and surgery.

As applied to jury deliberation, individual mock jurors could choose an avatar, sit in a jury box alongside other “players,” and, at the end of trial, retire to the jury room to deliberate with them. The “jury game” would simulate the courtroom environment, with lifelike characters (attorneys, judge, witnesses, fellow jurors, etc.) interacting directly with the participant. We envision a single player interacting with other computer-generated characters, much like a role-playing video game; but just as with computer-mediated deliberation, it might eventually be possible to have multiple participants deliberating in a VR jury simultaneously. In addition to being a more engaging research procedure, a VR jury paradigm could be used as a means of educating real jurors as part of the court’s juror orientation procedures—showing them what the task and environment will look like, introducing them to various court personnel, and so forth. At the rate VR technology is growing, along with its commercial, educational, clinical, and research applications, “jury game” technology could be available to jury researchers before we know it.

CONCLUSION

Although most extant jury research is, technically speaking, juror research, jury deliberations are a source of psychologically and legally rich data. Jury research has now been conducted for over one hundred years, dating back at least to the work of Hugo Münsterberg. As with research on individual jurors, methods for studying deliberating juries all have their pros and cons. Compared to Münsterberg, contemporary jury research methods are more realistic in some respects but less realistic in others. Although studies seldom involve deliberating groups, they overwhelmingly employ tasks that are informed by relevant legal issues and resemble situations real jurors would face. Future research in the

158. Id. at 4.
159. Id. at 5–7. These were the top five subject categories for both the entire study period (1990–2016) and the last five years, though the precise ranking varied slightly between the two.
160. We are grateful to Miko Wilford for showing us how this kind of technology could be applied to a legal decision-making task.
162. MÜNSTERBERG, SOCIAL SANITY, supra note 38, at 118, 121–22.
163. See, e.g., DEVINE, STATE OF THE SCIENCE, supra note 1, at 153. There is also some evidence that jury simulations have become more realistic in just the last twenty or so years, at least in terms of relying more on nonstudents as mock jurors and using nonwritten (i.e., audio, video, or live) trial materials. Bornstein, supra note 12, at 215–17.
field is likely to see a larger proportion of jury, as opposed to juror, studies as newer technologies make computer-mediated deliberation and other technologically driven methods (e.g., VR) more practical and affordable. These techniques solve some problems (e.g., the logistical issues involved in scheduling a six-person or larger group) while introducing others (e.g., the psychological effects of relative anonymity and greater spatial/social distance). Additional research is called for that examines the effects—both main and interactive—of new methodologies for studying juries. About the only thing we can know for certain is that the jury research of the future will continue to evolve.

164. A main effect would exist if, say, a fully computer-mediated jury were more likely to convict than a face-to-face jury. An example of an interactive effect would be if some other variable, like an angry outburst by one of the jurors, exerted differently in a computer-mediated versus a face-to-face deliberation. Interactions with methodological variables are typically more troubling, albeit more scientifically interesting, than possible main effects of those variables. Bornstein, supra note 12, at 217–18; Bornstein, Ecological Validity, supra note 103, at 88.