YALE LAW & POLICY REVIEW

The Effects of Comparable-Case Guidance on Awards for Pain and Suffering and Punitive Damages: Evidence from a Randomized Controlled Trial

Hillel J. Bavli[*] *& Reagan Mozer*[**]

Damage awards for pain and suffering and punitive damages are notoriously unpredictable. Courts provide minimal, if any, guidance to jurors determining these awards, and apply similarly minimal standards in reviewing them. Lawmakers have enacted crude measures, such as damage caps, aimed at curbing award unpredictability, while ignoring less drastic alternatives that involve guiding jurors with information regarding damage awards in comparable cases ("comparable-case guidance" or "prior-award information"). The primary objections to the latter approach are based on the argument that, because prior-award information uses information regarding awards in distinct cases, it introduces the possibility of biasing the award, or distorting the award size, even if prior-award information reduces the variability of awards. This paper responds to these objections. It reports and interprets the results of a large randomized controlled trial designed to test juror behavior in response to prior-award information and, specifically, to examine the effects of prior-award information on both variability and bias under a range of conditions related to the foregoing objections. We conclude that there is strong evidence that prior-award information improves the "accuracy" of awards—that it significantly reduces the variability of awards, and that any introduction of bias, or distortion of award size, is minor relative to its beneficial effect on variability. Furthermore, we conclude that there is

[*] Assistant Professor, SMU Dedman School of Law; Fellow, Harvard Institute for Quantitative Social Science.

[**] Ph.D. Candidate, Department of Statistics, Harvard University. This paper benefited immensely from the guidance of Professors Donald Rubin, Eric Posner, Michael Saks, Richard Lempert, David Rosenberg, Christopher Robertson, and James Greiner, and from support provided by the Institute for Quantitative Social Science and the Center for American Political Studies.

evidence that jurors respond to prior-award information as predicted in recent literature, and in line with the "optimal" use of such information; and that prior-award information may cause jurors to approach award determinations more thoughtfully or analytically.

I. INTRODUCTION

In civil jury trials, juries are frequently asked to determine awards for pain and suffering or punitive damages. But courts provide minimal, if any, guidance to jurors determining these awards, and apply similarly minimal standards in reviewing them. Consequently, these damage awards are notoriously unpredictable.[1] Courts have long expressed concern regarding

[1] *See* Payne v. Jones, 711 F.3d 85, 94 (2d Cir. 2013); Exxon Shipping Co. v. Baker, 554 U.S. 471, 499-503 (2008).

this "virtually unbridled discretion" of juries,[2] and the need to address this "standardless, unguided exercise of discretion by the trier of fact, reviewable . . . pursuant to no standard to guide the reviewing court either."[3] Indeed, award unpredictability can cause failures in deterrence and "corrective justice" objectives, harm to economies, high insurance premiums, and loss of faith in the legal system.[4]

Lawmakers have enacted crude measures, such as damage caps, aimed at curbing award unpredictability, while largely ignoring less drastic alternatives that involve guiding jurors with information regarding damage awards in comparable cases ("comparable-case guidance" or "prior-award information").[5] The primary objections to the latter approach are based on the argument that, because prior-award information uses information regarding awards in distinct cases, it introduces the possibility of biasing the award, or distorting the award size, even if prior-award information reduces the variability of awards.[6] This article responds to these objections. In particular, it reports and interprets the results of a large randomized controlled trial designed to test juror behavior in response to prior-award information and, specifically, to examine the effects of prior-award information on both variability and bias under a range of conditions related to the foregoing objections.

First, we examine whether prior-award information reduces award unpredictability. Some authors have argued that providing prior awards that are themselves subject to arbitrariness may only exacerbate the

[2] Geressy v. Digital Equip. Corp., 980 F. Supp. 640, 656 (E.D.N.Y. 1997) (quoting Leslie A. Rubin, Note, *Confronting a New Obstacle to Reproductive Choice: Encouraging the Development of RU-486 Through Reform of Products Liability Law*, 18 N.Y.U. REV. L. & SOC. CHANGE 131, 146 (1990-91)).

[3] Jutzi-Johnson v. United States, 263 F.3d 753, 759 (7th Cir. 2001); *see* Hillel J. Bavli, *The Logic of Comparable-Case Guidance in the Determination of Awards for Pain and Suffering and Punitive Damages*, 85 U. CIN. L. REV. 1, 5-8 (2017) [hereinafter *The Logic of CCG*].

[4] *See Payne*, 711 F.3d at 94; Oscar G. Chase, *Helping Jurors Determine Pain and Suffering Awards*, 23 HOFSTRA L. REV. 763, 768-69 (1995).

[5] *See The Logic of CCG*, *supra* note 3, at 8-12.

[6] *See, e.g.*, David A. Logan, *Juries, Judges, and the Politics of Tort Reform*, 83 U. CIN. L. REV. 903, 943-44 (2015) ("While [such an approach] would improve predictability, [it] would only be as good as the quality of the methodology for selecting which cases were factually similar enough to be included in the range"); *The Logic of CCG*, *supra* note 3, at 4-5.

problem of unpredictability.[7] We address this argument by testing the effect of prior-award information on the dispersion of awards under various conditions, including conditions related to the variability of the *prior* awards. We then confirm whether prior-award information has a biasing effect on damage awards—that is, whether it causes a distortion of award size—where the prior-award information arises from cases that are factually distinct from the present case. Then, to address the primary objections to prior-award information, we analyze the effect of prior-award information on variability *relative to* bias. Specifically, we develop a framework for examining the effect of prior-award information on "accuracy," which we define in terms of both variability and bias.

We conclude that there is strong evidence that prior-award information reduces the unpredictability of damage awards while also introducing the possibility of biasing the awards. Most importantly, we find strong evidence that prior-award information improves the accuracy of awards—that is, its beneficial effect on the dispersion of awards overwhelmingly dominates any distortion of award size. This occurs even when we design prior-award information specifically to introduce bias. In particular, we simulate conditions under which a court, for whatever reason, fails to align the facts of the current case with the facts of prior cases, thereby leading to an outlandish set of prior awards and ultimately to substantial bias. Even under these conditions, however, we find that any introduction of bias is minimal relative to the effects on variability.

Additionally, we examine a number of "behavioral" effects and find evidence that jurors respond to prior-award information as predicted in recent literature and in line with the "optimal" use of such information. We also find, based on our analysis of textual explanations provided by participants, that prior-award information may cause jurors to approach award determinations more thoughtfully or analytically.

In Part II, we provide a brief overview of the problem—the unpredictability of awards for pain and suffering and punitive damages—and describe a number of methods proposed to address it, including methods involving prior-award information. In Part III, we explain our methodology. In Part IV, we report and interpret our results. In Part V, we

[7]. *See, e.g.*, Mark Geistfeld, *Placing A Price on Pain and Suffering: A Method for Helping Juries Determine Tort Damages for Nonmonetary Injuries*, 83 CALIF. L. REV. 773, 792 (1995) ("If the system has been providing overly arbitrary pain-and-suffering awards, and if we have no method for determining the appropriate award in the first instance, why should we make prior awards the cornerstone of future awards? By doing so, we may ensure that like cases are treated alike in that all involve inappropriate damages awards.").

discuss limitations related to our methodology. Finally, in Part VI, we discuss certain implications of our analysis and we conclude.

II. Reducing Variability with Prior-Award Information

Awards for pain and suffering and punitive damages are notoriously unpredictable.[8] The Supreme Court and lower courts have repeatedly emphasized the importance of addressing the "inherent uncertainty of the trial process"[9] and the ruinous effects of such unpredictability.[10] Courts

[8] *See* Randall R. Bovbjerg et al., *Valuing Life and Limb in Tort: Scheduling "Pain and Suffering,"* 83 Nw. U. L. Rev. 908, 919-24 (1989) ("Tort law's traditional methods of computing damages for personal injury and death are under attack—and understandably so. Legal reformers have long argued that present law, when combined with jury discretion, inflates damage awards and creates problematic outcome variability. The open-ended and unpredictable nature of tort exposure has, in turn, threatened the liability insurance system that funds most tort compensation."); Chase, *supra* note 4, at 768-69 (citing studies); Shari Seidman Diamond et al., *Juror Judgments About Liability and Damages: Sources of Variability and Ways to Increase Consistency*, 48 DePaul L. Rev. 301, 317 (1998) (examining the "considerable variation in both juror and jury awards"); David W. Leebron, *Final Moments: Damages For Pain and Suffering Prior to Death*, 64 N.Y.U. L. Rev. 256, 259 (1989) ("The data . . . suggest that tort awards for even this relatively simple area [pain and suffering prior to death] vary significantly and that neither the specific facts of the case nor differing theoretical views of the functions of the awards can explain such variation."); Joni Hersch & W. Kip Viscusi, *Punitive Damages: How Judges and Juries Perform*, 33 J. Leg. Stud. 1, 1-10 (2004) (examining unpredictable awards for punitive damages); *see also The Logic of CCG*, *supra* note 3, at 1-5 (citing cases and literature). *But see* Theodore Eisenberg et al., *Variability in Punitive Damages: Empirically Assessing Exxon Shipping Co. v. Baker*, 166 J. Institutional & Theoretical Econ. 5 (finding that the data does "not support the unpredictability concern"); Yun-chien Chang et al., *Pain and Suffering Damages in Personal Injury Cases: An Empirical Study*, 14 J. Empirical Legal Stud. 199 (2017) ("[P]ain and suffering damages in Taiwan are to a large extent statistically and legally predictable."). *See generally* Neil Vidmar & Mirya Holman, *The Frequency, Predictability, and Proportionality of Jury Awards of Punitive Damages in State Courts in 2005: A New Audit*, 43 Suffolk U. L. Rev. 855 (2010).

[9] Exxon Shipping Co. v. Baker, 554 U.S. 471, 500-01 (2008) (quoting BMW of N. Am., Inc. v. Gore, 646 So.2d 619, 626 (Ala. 1994)).

[10] *Id.* at 499 ("The real problem, it seems, is the stark unpredictability of punitive awards"; "[t]hus, a penalty should be reasonably predictable in its

and commentators have proposed and implemented various methods of addressing the unpredictability of awards for pain and suffering and punitive damages, but none has prevailed as both adequate and appropriate.

A. Current Methods and Proposals

First, courts currently use the procedures of additur and remittitur—whereby a court that finds an award to be inadequate or excessive may order a new trial if the litigant harmed by the procedure does not agree to an increase (additur) or a reduction (remittitur) of the award.[11] But these devices are generally inadequate as tools for addressing variability. They

severity, so that even Justice Holmes's 'bad man' can look ahead with some ability to know what the stakes are in choosing one course of action or another. And when the bad man's counterparts turn up from time to time, the penalty scheme they face ought to threaten them with a fair probability of suffering in like degree when they wreak like damage." (citing Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 459 (1897))); BMW of N. Am., Inc. v. Gore, 517 U.S. 559, 574 (1996) ("Elementary notions of fairness enshrined in our constitutional jurisprudence dictate that a person receive fair notice not only of the conduct that will subject him to punishment, but also of the severity of the penalty that a State may impose"); Payne v. Jones, 711 F.3d 85, 94 (2d Cir. 2013) ("Apart from impairing the fairness, predictability and proportionality of the legal system, judgments awarding unreasonable amounts as damages impose harmful, burdensome costs on society."); Geressy v. Digital Equip. Corp., 980 F. Supp. 640, 656 (E.D.N.Y. 1997) (commenting on the "virtually unbridled discretion" of juries in deciding awards for pain and suffering); Chase, *supra* note 4, at 768-69 ("Variability is a problem primarily because it undermines the legal system's claim that like cases will be treated alike; the promise of equal justice under law is an important justification for our legal system. Variability is also claimed to create instrumental defects . . . ."); Bovbjerg et al., *supra* note 8, at 908 ("Determination of awards on an ad hoc and unpredictable basis, especially for 'non-economic' losses . . . tends to subvert the credibility of awards and hinder the efficient operation of the tort law's deterrence function"); *see also The Logic of CCG*, *supra* note 3, at 5-8 (citing relevant cases and literature).

[11] *See* David Baldus et al., *Improving Judicial Oversight of Jury Damages Assessments: A Proposal for the Comparative Additur/Remittitur Review of Awards for Nonpecuniary Harms and Punitive Damages*, 80 IOWA L. REV. 1109, 1118-20 (1995).

are applied inconsistently and with minimal standards;[12] they are used to address only the most extreme awards rather than variability generally; and regular use of such methods, and the replacement of jury determinations with the discretion of the court, would arguably raise significant constitutional problems and would be inconsistent with norms of tort law.[13] Additionally, these methods serve as a band-aid rather than addressing the underlying problem—that juries receive insufficient guidance for assessing awards for pain and suffering and punitive damages.[14]

Second, numerous jurisdictions have imposed damage caps to address extreme awards. Legislatures have enacted damage caps for certain types of awards, such as punitive damages, or for damage awards generally.[15] However, damage caps address only extreme cases and only excessive awards.[16] Moreover, capping awards wholesale without regard for the individual circumstances of a case gives rise to fairness and proportionality concerns and can harm the deterrence objectives of tort law and disincentivize beneficial lawsuits.[17] It may cause constitutional concerns as well.[18]

Third, a number of commentators have proposed using awards in comparable cases as guidance for award determinations. These methods

[12] Jutzi-Johnson v. United States, 263 F.3d 753, 759 (7th Cir. 2001) ("[Most courts] treat the determination of how much damages for pain and suffering to award as a standardless, unguided exercise of discretion by the trier of fact, reviewable for abuse of discretion pursuant to no standard to guide the reviewing court either.").

[13] *See* Baldus et al., *supra* note 11, at 1118-27; *The Logic of CCG*, *supra* note 3, at 8-9. *But see* CASS SUNSTEIN ET AL., PUNITIVE DAMAGES: HOW JURIES DECIDE 248-52 (2002) (arguing for a larger judicial role in determining punitive damages).

[14] Note that *The Logic of CCG*, *supra* note 3, proposes the use of prior-award information *in addition* to the procedures of additur and remittitur—not in place of them.

[15] *See* Joseph Sanders, *Why Do Proposals Designed to Control Variability in General Damages (Generally) Fall on Deaf Ears? (and Why This Is Too Bad)*, 55 DEPAUL L. REV. 489, 510 (2006).

[16] *The Logic of CCG*, *supra* note 3, at 9.

[17] *See* Sanders, *supra* note 15, at 509-11; *The Logic of CCG*, *supra* note 3, at 9.

[18] Kathryn Zeiler, *Turning from Damage Caps to Information Disclosure: An Alternative to Tort Reform*, 5 YALE J. HEALTH POL'Y L. & ETHICS 385, 387 (2005).

have been proposed in various forms. Some have focused on the use of comparable cases to develop a more principled approach to judicial review of jury awards for excessiveness.[19] Although these methods would improve the standards underlying a court's review, they suffer from many of the same problems that apply to the additur and remittitur procedures. For example, they address only extreme awards, and regular replacement of the jury's discretion with that of the court arguably gives rise to constitutional concerns and is inconsistent with norms of tort law. Similar issues arise from methods that involve binding the trier of fact to a particular award or range of awards, or that predetermine a schedule of awards in advance of a case. These methods have been criticized as replacing the jury's discretion with that of the court or a legislative body altogether removed from the subject case.[20]

Some recommendations, however, involve "comparability analysis," whereby a court identifies comparable cases, provides the trier of fact with information regarding the awards in these cases, and instructs the trier of fact to arrive at a damages award based on the evidence, using the comparable-case information as guidance.[21] These recommendations are based on studies demonstrating that they are effective methods of controlling outlying awards and variability generally, even when the information is provided as non-binding guidance.[22]

B. Prior-Award Information

In this paper, we focus on the specific type of comparability analysis called "comparable-case guidance" (CCG) or "prior-award information," described in *The Logic of CCG* as information derived from prior

[19] *See, e.g.*, Baldus et al., *supra* note 11.

[20] *The Logic of CCG*, *supra* note 3, at 10; *see* Logan, *supra* note 6, at 942-43 ("Such an approach would streamline litigation and greatly limit, if not eliminate, the concerns with variability and fairness that the current practice risks by treating like cases differently. However, this approach is fatally flawed because it eviscerates the various contributions that juries make to the civil justice system. Moreover, this approach is fundamentally inconsistent with the basic tort principle that each victim is entitled to an award tailored to his or her circumstances, set by a lay jury." (citing RESTATEMENT (SECOND) OF TORTS § 901 cmt. a (AM. LAW INST. 1979))).

[21] *The Logic of CCG*, *supra* note 3, at 3.

[22] *See* Michael J. Saks et al., *Reducing Variability in Civil Jury Awards*, 21 L. & HUM. BEHAV. 243, 249-55 (1997).

"comparable" cases considered by the trier of fact as guidance (as opposed to a binding range) for determining damage awards.[23]

Numerous courts and commentators have called for the use of prior-award information (in some form or other) to guide jury determinations. Consider, for example, the case *Jutzi-Johnson v. United States*, which involved an appeal from an award for pain and suffering resulting from a bench trial.[24] In that case, Judge Richard Posner commented on the "acute" problem of "figuring out how to value pain and suffering."[25] According to Judge Posner, notwithstanding "[v]arious solutions, none wholly satisfactory, [that] have been suggested," "[m]ost courts do not follow any of these approaches. Instead, they treat the determination of how much damages for pain and suffering to award as a standardless, unguided exercise of discretion by the trier of fact, reviewable for abuse of discretion pursuant to no standard to guide the reviewing court either."[26] He concluded that "[t]o minimize the arbitrary variance in awards bound to result from such a throw-up-the-hands approach, the trier of fact should, as is done routinely in England . . . be informed of the amounts of pain and suffering damages awarded in similar cases."[27] He continued: "And when the trier of fact is a judge, he should be required as part of his Rule 52(a) obligation to set forth in his opinion the damages awards that he considered comparable," noting that courts "make such comparisons routinely in reviewing pain and suffering awards," and remarking that "[i]t would be a wise practice to follow at the trial level as well."[28]

[23] *The Logic of CCG*, *supra* note 3, at 4.

[24] 263 F.3d 753 (7th Cir. 2001).

[25] *Id.* at 758.

[26] *Id.* at 758-59.

[27] *Id.* at 759(citations omitted).

[28] *Id.* (citations omitted); *see also* Roselle Wissler et al., *Decisionmaking About General Damages: A Comparison of Jurors, Judges, and Lawyers*, 98 MICH. L. REV. 751, 816-17 (1999) (discussing "reforms consistent with the available data" and suggesting that "[a]nother powerful yet modest reform would be to pool jury awards made for similar injuries, and to present these cases and their award distributions to juries for guidance in reaching their general damages awards and to judges for conducting their additur/remittitur reviews"); Chase, *supra* note 4, at 775, 777-90 (discussing recommendation by the ABA Action Commission to Improve the Tort Liability System to establish "'tort award commissions' . . . to gather and report information that would be useful in 'the framing of jury instructions, the exercise of the power of additur and remittitur, and the process of settling cases,'" and proposing

However, calls for such methods have generally failed, in part due to objections based on their reliance on a court's ability to identify an appropriate set of "comparable" cases.[29] First, some authors have questioned whether prior-award information can increase predictability, notwithstanding the unpredictability of the prior awards themselves.[30] Second, and most prominently, commentators have argued that, even assuming that prior-award information reduces unpredictability, its reliance on awards arising from *distinct* cases introduces the possibility of biasing the award, or distorting the size of the award. The fundamental issue is therefore not simply whether prior-award information reduces variability. It is whether it reduces variability *relative to* any introduction of bias.[31]

Therefore, we address these objections by examining the effect of prior-award information on "error"—or, inversely, "accuracy"—which we define in terms of both variability and bias.[32] Conceptually, accuracy can be understood as reflecting the idea that a damages award can be "better" or "worse" based on societal objectives and norms.[33] For purposes of this paper, we use the term in a specific way: we improve the accuracy of an award if we reduce the unpredictability of the award without introducing "too much" bias. In Part III, we provide formal definitions of accuracy, error, variability, and bias within the context of our experiment, and we

method involving charts providing nonbinding guidance "to allow comparison with roughly similar cases in which plaintiffs' verdicts were recovered") (quoting *Report of the Action Commission to Improve the Tort Liability System*, AM. BAR ASSOC. 10-15 (1987)); Logan, *supra* note 6, at 939-44 (discussing proposals); Sanders, *supra* note 15, at 496-507 (discussing proposals and studies).

[29] *See The Logic of CCG*, *supra* note 3, at 4-5.

[30] *See* Geistfeld, *supra* note 7, at 792; Peter H. Schuck, *Scheduled Damages and Insurance Contracts for Future Services: A Comment on Blumstein, Bovbjerg, and Sloan*, 8 YALE J. REG. 213, 218 (1991) ("[B]y using earlier awards as the foundation for their new system of damages scheduling, they impound and then compound what they themselves characterize as the distortions of the past, thereby projecting those distortions into the future."); *The Logic of CCG*, *supra* note 3, at 4-5.

[31] *The Logic of CCG*, *supra* note 3, at 5.

[32] *Id.* at 13.

[33] Courts arguably recognize this when, for example, they use tools such as additur and remittitur.

explain how much bias is "too much," so as to reduce accuracy. Here, however, let us examine these notions in further detail conceptually.

Previous literature has modeled a damages award as an estimation problem in which an actual award serves as an *estimate* of a "correct" award that would reflect complete information regarding the law and the facts of a claim. In this approach, the error associated with the estimate represents the distance between the actual award and the correct award.[34] In turn, according to this model, error can be deconstructed into variance and bias.[35] Variance is a measure of dispersion around a mean, or an "expected," award value. Bias, meanwhile, is a measure of the difference between the expected award and the correct award. "If [the estimation process] is 'unbiased,' then it will generate the correct value *on average*. If it is 'biased,' then it will generate the incorrect value on average, and the 'bias' reflects the distance between the value the [the estimation process] generates on average and the correct value."[36] Unbiasedness is generally a desirable feature of an estimation process, but it does not imply the

[34] *The Logic of CCG*, *supra* note 3, at 12 (citing Hillel J. Bavli, *Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation*, 14 L., PROBABILITY & RISK 67, 74-78 (2015) [hereinafter *Aggregating for Accuracy*]). Note that we can similarly define a *distribution* of "correct" awards reflecting, for example, uncertainty regarding the law. *See* Hillel J. Bavli, *Sampling and Reliability in Class Action Litigation*, 2016 CARD. L. REV. DE NOVO 207, 209 n.16 (2016). *The Logic of CCG* follows previous literature that characterizes a "correct" award as "the mean of the population of possible awards that would emerge from adjudicating the case repeatedly under various conditions (e.g., before different judges and juries, by different attorneys, with different permutations of facts, etc.)." *The Logic of CCG*, *supra* note 3, at 13 (citing *Aggregating for Accuracy*, *supra* note 34, at 74-78; Michael J. Saks & Peter David Blanck, *Justice Improved: The Unrecognized Benefits of Aggregation and Sampling in the Trial of Mass Torts*, 44 STAN. L. REV. 815, 833-34 (1992)). This characterization is intended to capture various interpretations of the law, the facts of the case, norms, etc. *The Logic of CCG*, *supra* note 3, at 13 n.51. There are in fact many reasonable definitions for the "correct" award in a case. Even using the conceptualization above, other measures of central tendency—such as the median—may be used, depending on our beliefs regarding the best way to characterize the "correct" award (e.g., whether we want to capture information regarding extreme values, etc.). We adopt this characterization; but, as discussed, it is for convenience rather than necessity. *See generally* RONALD DWORKIN, LAW'S EMPIRE (1986).

[35] *The Logic of CCG*, *supra* note 3, at 12-15.

[36] *Id.* at 14.

absence of error: the process may produce highly dispersed values around a correct value.[37] For example:

> [I]f the correct punitive damages value in [a] case . . . is $100,000, then repetitions of an unbiased adjudication may generate estimate values (i.e., damage awards) of $0, $50,000, $150,000, and $200,000, which are indeed centered at the correct value of $100,000; however, the awards are highly dispersed around $100,000. We would, for example, prefer that repeated adjudications generate the values $90,000, $95,000, $105,000, and $110,000; or even better, $100,000, $100,000, $100,000, and $100,000.[38]

Therefore, variance, a measure of such dispersion, is also a crucial component of error.[39] Indeed, some circumstances may involve a tradeoff between bias and variance, where it is necessary to introduce the possibility of some bias in order to achieve substantial reductions in variance and, on balance, significant improvements in accuracy. For example, if the correct award in a case is $100,000, it may be preferable to have an estimation process that generates the values $90,000, $93,000, $97,000, and $100,000 rather than $0, $50,000, $150,000, and $200,000, even though the latter set is unbiased and therefore centered at the correct award while the former set is biased and therefore centered at the incorrect award. Specifically, the former set may be preferable because, while it involves some bias, the bias is modest (the mean of the four values is $95,000—only $5,000 off from the correct award) and the variance is far less than that associated with the latter set of awards.

It has been argued that applying prior-award information to improve the accuracy of damage awards involves this type of tradeoff.[40] Specifically, it has been argued that "CCG improves the accuracy of awards for pain and suffering and punitive damages—award types that suffer from particularly high degrees of variability—by facilitating a balance between minimizing variability and introducing the possibility of bias."[41] *The Logic of CCG* examined this tradeoff theoretically. The purpose of the current article is to test it experimentally.

[37] *Id.*

[38] *Id.*

[39] *Id.*

[40] *Id.* at 15-24.

[41] *Id.* at 19.

Thus, in terms of bias and variance, prior-award information is hypothesized to cause a reduction in variance by providing jurors with a context to guide their decision, or some form of "anchor,"[42] but at the inherent cost of introducing the risk of bias by providing award information from cases that are not identical to the subject case.[43]

Importantly, for purposes of this paper, it is unnecessary to assume the existence of a correct award. We are concerned only with whether prior-award information improves or harms the accuracy of an award relative to what the award would be in the absence of prior-award information. In this sense, rather than requiring a correct award, our argument relies on two assumptions: that reducing the variability of an award is good (i.e., improves the accuracy of an award) and introducing bias is bad (i.e., harms the accuracy of an award). Defining a correct award is useful for purposes of simplifying terminology, and to remain consistent with terminology in previous literature. Further, our definition of a correct award promotes a conservative analysis. Specifically, as in previous literature, we define the correct award associated with a claim as the average of repeated adjudications of a claim without prior-award information.[44] For example, we assume that if a court had unlimited resources, it would prefer to adjudicate a claim numerous times (before numerous juries, etc.) and use the average adjudication as the ultimate damages award. This allows for the foregoing assumptions—that reducing variance is good and introducing bias is bad—by assuming that the award-generating process is initially unbiased. Again, this assumption of unbiasedness is not necessary for our results; indeed, it is unlikely that damage awards are entirely unbiased. Rather, we include this assumption for purposes of convenience and to promote conservative estimates by ensuring that any bias introduced by prior-award information counts as "bad" in the sense of reducing accuracy (in actuality, if an award is in fact biased, introducing bias through prior-award information could improve accuracy by negating the preexisting bias).

Earlier articles have used statistical theory and modeling assumptions—in particular, the optimal use of prior-award information, given the variability of the prior awards and the random variation of the subject award—to argue that prior-award information would improve the accuracy of awards for pain and suffering and punitive damages under a

[42] Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1128-30 (1974).

[43] *The Logic of CCG*, *supra* note 3, at 17-19.

[44] *Id.* at 12-15

robust set of conditions regarding the actual comparability of the cases underlying the prior awards relative to the subject case. These "comparability" conditions are defined in terms of both the "misalignment" of the prior cases with the present case—i.e., the difference between the mean of the correct awards in the prior cases and the correct award in the subject case—and the substantive (or factual) variability of the prior cases, reflected in the variance of the prior awards.[45] *The Logic of CCG* discussed methods for identifying comparable cases and extracting comparable-case information for consideration by a jury. This process does not require a reinvention of the wheel: courts have used comparability analyses in various contexts and these contexts provide good guidance for developing comparable-case information to guide juries in their determinations of awards for pain and suffering and punitive damages.[46] For example, courts have used comparability analysis for purposes of additur and remittitur, on appeal in reviewing awards for excessiveness, and in various substantive contexts, such as challenges to compensation in the takings context.[47]

A detailed analysis of methods for developing comparable-case information is beyond the scope of this article. But as discussed in *The Logic of CCG*, courts should be concerned with balancing three factors when developing such information: "the [factual] alignment of the prior cases with the subject case, the [factual] breadth of the prior cases, and the number of prior cases."[48] As courts have done in other contexts, a court might provide guidance to litigation parties for identifying comparable cases and then "select[] a final set of cases from a pool identified by the litigants and the court."[49] Further, courts can implicitly involve jurors in the selection process by including fact summaries of the selected prior cases in the presentation of comparable-case information.[50] Of course, methods for identifying prior cases and distilling information for

[45] *See id.* at 15-24; Hillel J. Bavli & Yang Chen, *Shrinkage Estimation in the Adjudication of Civil Damage Claims,* 13 REV. L. & ECON. (2017), http://doi.org/10.1515/rle-2015-0010.

[46] *See The Logic of CCG*, *supra* note 3, at 24-28.

[47] *Id.* at 28-31.

[48] *Id.* at 24.

[49] *Id.* at 26.

[50] *Id.*

presentation to a jury must be balanced against accompanying litigation costs.[51]

Ultimately, these earlier articles on comparable-case guidance have concluded that the accuracy benefits of prior-award information are not very sensitive to a court's ability to identify a set of comparable cases. They conclude that, absent very substantial error in selecting prior cases, any introduction of bias caused by prior-award information would be minimal relative to the beneficial reduction in the random variation of awards, and that, given a reasonable method for identifying prior cases, providing jurors with prior-award information would improve the accuracy of award determinations.[52]

Thus, our aim in this paper is to test whether these effects hold empirically, and whether they result from the mechanisms described in the foregoing models. Our study differs from previous studies in that we aim particularly to address the primary objections to these methods in the literature by focusing explicitly on the effects of prior-award information on the accuracy of awards under a range of conditions related to the selection of prior awards.[53] In the following part, we explain our framework for analyzing the effects of prior-award information on variability and bias and for analyzing these effects together and relative to each other under varied conditions.

[51] *Id.* at 26-28.

[52] *See id.* at 12-24; Bavli & Chen, *supra* note 45, at 17-24.

[53] There have been numerous studies examining the unpredictability of award determinations and methods of addressing it. *See, e.g.*, Baldus et al., *supra* note 11; Bovbjerg et al., *supra* note 8; Diamond et al., *supra* note 8; Leebron, *supra* note 8; Saks et al., *supra* note 22; Catherine M. Sharkey, *Unintended Consequences of Medical Malpractice Damages Caps*, 80 N.Y.U. L. REV. 391 (2005); *see also* Chase, *supra* note 4; Sanders, *supra* note 15, at 496-507 (summarizing proposals and studies). The experiment by Professor Saks et al., *supra* note 22, is most closely related to the current study. That study investigated the effects of prior awards in various forms, as well as a "cap condition," on the amounts awarded for pain and suffering in personal injury cases—and specifically, with respect to award variability and "distortions" in the amounts awarded. *See id.* at 246-47. The authors found that "interval," "average-plus-interval," and "examples" conditions—involving prior awards that, in the terms of the current paper, were "aligned" with the subject cases based on pilot studies—caused a reduction in variability while distorting award amounts minimally or not at all. *See id.* at 246. They also found that the "cap" condition performed poorly and sometimes increased the variability and size of awards. *See id.* at 253.

III. Methodology

To test the effects of prior-award information on the spread, magnitude, and accuracy of awards for pain and suffering and punitive damages we conducted a randomized controlled trial (RCT) using a factorial design and the potential-outcomes framework.[54] Specifically, we exposed each participant to a "treatment combination" (or "treatment condition") arising from a set of "factors" reflecting the prior-award conditions of interest. In each treatment condition, we set each of these factors to a certain value, or "level."[55] We then drew causal conclusions based on inferences about what an outcome would be under exposure to alternative treatment conditions.[56] In this part, we describe our methodology, including the details of our experimental design and our analysis.

A. RCT Setup and Administration

The study involved two legs, a primary leg (Leg I) and a secondary leg (Leg II). These should be understood as two separate RCTs. Parts III and IV of this article refer to Leg I unless stated otherwise. We administered Leg II specifically to address questions that arose from Leg I; accordingly, we refer to those results as related questions arise.

We used Amazon's Mechanical Turk to recruit and administer the experiment to approximately 5,500 participants in Leg I and 2,500 participants in Leg II. We restricted participation to U.S. citizens who were eighteen or older and English-speaking, the baseline eligibility requirements for serving on a U.S. jury.[57] Participants enrolled in the

[54] *See* Tirthankar Dasgupta, Natesh S. Pillai, & Donald B. Rubin, *Causal Inference from 2^K Factorial Designs by Using Potential Outcomes*, 77 J. ROYAL STAT. SOC. SERIES B (STAT. METHODOLOGY) 727, 727 (2015) [hereinafter *2^K Factorial Designs*]; Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. EDUC. PSYCHOL. 688 (1974). We use the notation and general framework established in *2^K Factorial Designs* but extend the notation to accommodate the additional levels present in this experiment.

[55] *See generally 2^K Factorial Designs*, *supra* note 54.

[56] *See generally* Rubin, *supra* note 54, at 689-90; C. F. JEFF WU & MICHAEL S. HAMADA, EXPERIMENTS: PLANNING, ANALYSIS, AND OPTIMIZATION (2d ed. 2009).

[57] Mechanical Turk requires that all participants be eighteen or older and English-speaking. We attempted to enforce the U.S. citizenship requirement

experiment sequentially and were randomly assigned to receive one of twenty-two treatment conditions, which determined what prior-award information, if any, a participant received as guidance for her response.[58] For each treatment condition, participants were provided a fact pattern and jury instructions that included guidance, specific to the assigned treatment condition, for determining a damages award. Participants were then asked to determine a damages award, provide an optional explanation, and provide certain demographic information.[59]

Each treatment condition contained one of two fact patterns. Conditions with a punitive damages fact pattern contained the following paragraph:

> You are a juror in a trial in which a car manufacturer concealed its knowledge of a defect in its car's airbag system. As a result of the defect, the airbags would fail to deploy in a small proportion of frontal collisions. The lawsuit was brought by a driver, Andrew, who suffered severe brain injury from a frontal collision caused by ice on the road. He now lives with headaches, blurred vision, speech impairment, and memory loss. At trial, it was established that, as a result of the defect, the airbags failed to deploy. It was also established that, had the airbags deployed properly, Andrew's injuries would have been avoided.

by requiring that all participants have an IP address located within the United States. Additionally, we asked that potential participants refrain from participating if they failed to meet the requirements.

[58] We used a "dynamic randomization scheme" to ensure roughly equal sample sizes across treatment groups. This scheme maintains the properties of classical randomization, assuming the participants enroll randomly in the sense that one individual's enrollment does not affect another's subsequent enrollment other than through the restriction on the number of units in each treatment condition.

[59] Participants received $0.20 to $0.30 for their participation in the survey. Participants were also given the opportunity to provide feedback regarding the survey. We used two measures to confirm that participants were attentive rather than completing the survey arbitrarily. First, we examined the amount of time that participants took to complete the survey, which comported with our expectations; and we examined the proportion of participants who provided optional explanations, which was high—about 75-80%. While we used explanations as a signal for attentiveness, we did not interpret the absence of an explanation as a signal of inattentiveness.

Conditions with a pain and suffering fact pattern contained a different paragraph:

> You are a juror in a trial in which a company intentionally disposed of its industrial waste by regularly dumping it into a local river rather than having the expense of disposing it properly. The lawsuit was brought by Emma, a 29-year-old woman whose drinking water was affected by the improper disposal and who developed a rare cancer as a result. Three years before her diagnosis, Emma married her college boyfriend. She and her husband now have a two-year-old daughter. Emma has undergone multiple surgeries and months of chemotherapy and radiation therapy, but doctors have recently informed her that the cancer has spread and that her likelihood of survival beyond six months is very low. Since her diagnosis one year ago, she has suffered from regular pain, nausea, fatigue, and disfigurement, and her organs have recently begun to fail.

All conditions also included a paragraph-long jury instruction based on real-world instructions for determining punitive damages[60] or a damages

[60] For the punitive damages scenario, stimuli included the following paragraph:

> The judge has asked you to determine a "punitive damages" award. He informs you that, through a separate proceeding, Andrew has already been compensated for his injuries, including his medical expenses and his pain and suffering. The judge instructs you that your role now is to determine a "punitive damages" award. He explains that "punitive damages are damages awarded not to compensate the plaintiff for any injury but to punish the defendant for outrageous conduct and to deter the defendant and others from similar conduct in the future. You are not required to award punitive damages, and you may award such damages only if you find that the defendant's conduct was in fact outrageous." The judge emphasizes that "there is no exact standard for determining punitive damages. You should decide on an amount that you find necessary for achieving the objectives described above. You should consider the degree of reprehensibility of the defendant's misconduct and the actual or potential harm suffered by the plaintiff."

See JOHN J. KIRCHER & CHRISTINE M. WISEMAN, 1 PUNITIVE DAMAGES: LAW AND PRACTICE 2d § 11:8 (2018 ed.) (containing pattern jury instructions adapted for the survey).

award for pain and suffering.[61] For the control arms of the experiment, the description ended and participants were asked to determine an award for punitive damages or pain and suffering (depending on the participant's assigned fact pattern). Participants assigned to one of twenty active treatment conditions, however, were additionally provided certain information regarding awards in prior "comparable" cases. For example, they may have been provided with a paragraph similar to the following:

> Additionally, the judge informs you that in five previous similar cases juries have determined awards for [punitive damages or pain and suffering] in the amounts of [$__________, $__________, $__________, $__________, and $__________]. The judge indicates that this information regarding prior awards is intended as guidance only, and that you may use (or not use) the information as you see appropriate.

The form and substance of the numerical prior-award values were based on the treatment condition to which a participant was randomly assigned—and particularly, based on the levels of *scenario*, *bias*, *variability*, and *form* associated with that treatment condition. These values are discussed in the following subsection.

[61] For the pain and suffering scenario, stimuli included the following paragraph:

> The judge has asked you to determine a suitable damages award for Emma's pain and suffering (past and future) and her loss of capacity for enjoyment of life. He informs you that, through a separate proceeding, Emma has already been compensated for her monetary costs, such as past and future medical expenses. The judge instructs you that your role now is to determine an award for Emma's physical and mental pain and suffering (past and future) and her loss of capacity for enjoyment of life. The judge emphasizes that "no evidence of the value of intangible things, such as mental or physical pain and suffering, has been or need be introduced. You are not trying to determine value, but an amount that will fairly compensate the plaintiff for the damages she has suffered. There is no exact standard for fixing the compensation to be awarded for these elements of damage." You should use your judgment to decide a fair amount.

See PATTERN CIV. JURY INSTR. 5TH CIR. § 15.3 (COMMITTEE ON PATTERN JURY INSTRUCTIONS, DISTRICT JUDGES ASSOCIATION, FIFTH CIRCUIT 2014) (containing pattern jury instructions adapted for the survey).

B. Treatment Conditions

As discussed, the two control conditions involved no prior-award information and involved either a punitive damages scenario or a pain and suffering scenario.[62] Each of the twenty active treatment conditions involved a combination of four experimental factors: scenario (two levels), form (three levels), bias (two levels), and variability (two levels), where scenario was set to either *punitive damages* or *pain and suffering*, form was set to either *average*, *list*, or *range*, bias was set to either *unbiased* or *biased*, and variability was set to either *low variability* or *high variability*. Therefore, each treatment condition is characterized by the level to which each factor is set, and the treatment condition to which a participant was randomly assigned determined the particular stimulus he or she received. For example, a participant assigned to the treatment condition that involves [scenario = pain and suffering, form = list, bias = unbiased, and variability = high variability] received a stimulus that asked the participant to determine an award for pain and suffering, and provided unbiased high-variability prior-award information in the form of a list.[63]

To determine the numerical values that would define each active treatment condition, we conducted a pilot study (n=400) that substantively replicated the control conditions of the main study, and we used the distributions of the award determinations in the pilot study as a reference for defining levels of bias and variability. For each level of scenario, we used the median of award amounts in the corresponding pilot sample as the unbiased average[64] and the 30th percentile of award

[62] For purposes of the factorial design, the control conditions can together be viewed as a separate factorial experiment.

[63] By design, certain levels of certain factors are incompatible with certain levels of other factors. In particular, the variability factor is not applicable when prior-award information is presented in the average form. Also, as previously stated, for control conditions, only the scenario factor is applicable.

[64] Implicit in our terminology is the existence of a correct award for each scenario. *See supra* Section II.B. We defined the correct award consistently with previous literature, as the mean of the distribution of awards that would result from infinitely repeated adjudications (of the control condition) under various conditions. *See The Logic of CCG*, *supra* note 3, at 12-13; *supra* note 34. For a number of reasons, however, the median is likely to serve as a better estimator than the mean. For example, the mean would be too heavily influenced by a few extreme outliers in the relatively small control samples (relative to infinitely repeated adjudications). Also, extreme

amounts as the biased average.[65] Additionally, for each level of scenario, we used the 15th, 25th, 50th, 75th, and 85th percentiles of award amounts in the corresponding pilot sample as the unbiased high-variability list, and the 40th, 45th, 50th, 55th, and 60th percentiles as the unbiased low-variability list. Finally, for each level of scenario, we used the 15th and 85th percentiles of the award amounts from the corresponding pilot sample as the unbiased high-variability range, and the 40th and 60th percentiles as the unbiased low-variability range.

To define high-variability and low-variability prior-award information in the biased treatment conditions, we calculated a "bias ratio" as the ratio of the 30th percentile to the median within each level of scenario, and then multiplied the values determined for the unbiased treatment conditions by the corresponding bias ratio. Table A lists the prior-award values that we provided to participants in the main study in each active treatment condition.

outliers are far less likely in real-world trials involving actual juries and presided over by judges. *See infra* Part V. In any event, our conclusions would not be weakened were we to analyze the effects of interest with reference to the results in the control condition directly, without reference to a correct award. For example, without taking a position on how to characterize a correct award, we could investigate the effect of a particular treatment condition on magnitude, relative to the magnitude of the award in the control condition. After all, prior-award information is intended to improve accuracy by reducing error from dispersion substantially more than it adds error from distortion—whatever the initial "distortion" may be. As discussed *supra* Section II.B, by referencing a correct award, we simply assume that the award determination is initially undistorted, or unbiased; and we assume that any distortion of award size is harmful. To be sure, if the initial bias were sufficiently extreme, we may, under certain conditions, not want to reduce variability around the central award. But such a problem would require extreme initial bias, and there is no reason to assume that such a bias exists (even if, for our purposes, there were cause to assume some source of bias in the first place). Furthermore, such an assumption may be inconsistent with the position of the courts that reducing random variation, without more (e.g., without causing bias), is beneficial. In any event, for purposes of clarity and consistency with previous literature, and to highlight the assumption that any introduction of bias is harmful to accuracy, we defined the correct awards as above, and estimated them using the medians of the control group awards.

[65] We defined "biased" conditions using the 30th percentile to reflect very substantial, but not entirely unrealistic, court "error" in determining a set of prior awards. In Leg II, we used an alternative definition. *See infra* Section III.E.

Table A. Prior-Award Information Provided in Active Treatment Conditions

Scenario	Bias and Variability	Form	Prior-Award Information[66]
Pain and Suffering	Unbiased Low Variability	Average	$2m
		Range	$1.12m to $4.7m
		List	$1.12m, $2m, $2m, $3m, $4.7m
	Unbiased High Variability	Average	$2m
		Range	$200k to $15m
		List	$200k, $500k, $2m, $10m, $15m
	Biased Low Variability	Average	$1m
		Range	$560k to $2.35m
		List	$560k, $1m, $1m, $1.5m, $2.35m
	Biased High Variability	Average	$1m
		Range	$100k to $7.5m
		List	$100k, $250k, $1m, $5m, $7.5m
Punitive Damages	Unbiased Low Variability	Average	$1m
		Range	$500k to $1m
		List	$500k, $500k, $1m, $1m, $1m
	Unbiased High Variability	Average	$ 1m
		Range	$10k to $10m
		List	$10k, $52.5k, $1m, $5m, $10m
	Biased Low Variability	Average	$100k
		Range	$50k to $100k
		List	$50k, $50k, $100k, $100k, $100k
	Biased High Variability	Average	$100k
		Range	$1k to $1m
		List	$1k, $5.25k, $100k, $500k, $1m

C. Assessing and Correcting Covariate Balance

We collected data on nine covariates—age, sex, ethnicity, education, employment status, marital status, household income, residential community type, and political affiliation—upon receiving each participant's award determination and optional explanation. We used these data, as well as metadata regarding participant enrollment times, to test whether our randomization achieved reasonable covariate balance

[66] We use "m" to denote millions and "k" to denote thousands.

across treatment conditions.[67] Substantial imbalances could prevent us from knowing whether any observed effects are attributable to the intervention or to the imbalance. Therefore, substantial imbalances should be corrected prior to proceeding to the analysis phase of an experiment.[68]

To assess covariate balance across the twenty-two treatment groups, we used Pearson's Chi-squared test, which revealed no statistically "significant" differences among the twenty-two treatment groups.[69] We also performed pairwise comparisons to evaluate differences in frequency distributions on each covariate for each of the $\begin{pmatrix} 22 \\ 2 \end{pmatrix} = 231$ pairs of distinct treatment groups using Fisher's exact test and the "Benjamini-Hochberg procedure" to control the false discovery rate (FDR) to correct for multiple comparisons.[70] Again, we found no significant differences between pairs of treatment groups on any of the eleven variables tested.

[67] We used metadata regarding participant enrollment times to construct variables reflecting the time of day and day of the week of participant enrollment to check for systematic differences in participants based on enrollment times that may influence overall covariate balance and to validate our assumption that participants enrolled in the study randomly.

[68] To avoid data "dredging" and ensure that covariate balance would be corrected in an objective manner, we finalized procedures for assessing covariate balance and for addressing any imbalances prior to accessing any data, and we analyzed the covariate data in a so-called secondary design phase in which we removed all outcome data and considered only covariate data.

[69] We followed the common practice of combining, where sensible, neighboring levels of the covariate contingency table to ensure that there were five or more (expected) observations in each cell of the table. *See generally* William G. Cochran, *The $\chi 2$ Test of Goodness of Fit*, 23 ANNALS MATHEMATICAL STAT. 315 (1952). For example, for age, we aggregated levels for ages above sixty-one to create one "Over 61" level; for ethnicity*,* participants who marked "American Indian or Alaska Native" were aggregated with participants marking the value, "Other." Additionally, for the covariate sex, we observed twenty participants who marked the response "Other" (averaging one or fewer observations per treatment group). Because there was no clear way to recode these observations into either the "Male" or "Female" levels, we excluded these participants from subsequent analyses. We then performed post-hoc analyses to evaluate any impact of excluding them and concluded that there was no such impact.

[70] *See* Yoav Benjamini & Yosi Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, 57 J. ROYAL STAT. SOC.

In summary, our analysis of the covariate data indicated that, by the observed randomization, we achieved suitable balance on all demographic variables and enrollment-time variables across the twenty-two treatment groups. Therefore, we proceeded to the analysis phase of the experiment without applying any corrections or adjustments to improve balance.

D. Causal Inference for Factorial Effects

The analysis of data from factorial experiments often relies on a generalized linear model framework (i.e., analysis of variance (ANOVA)). However, as discussed in *2^K Factorial Designs*, these approaches have drawbacks that can impede the ability to make causal conclusions about the experimental factors.[71] We therefore based our analyses and estimation of causal effects on the potential outcomes framework of Jerzy Neyman,[72] often referred to as the Rubin Causal Model (RCM).[73] We followed the basic notation and philosophy of estimation in *2^K Factorial Designs,* which developed a theoretical framework for causal inference from factorial designs using the potential outcomes model.[74]

SERIES B (STAT. METHODOLOGY) 289 (1995). We also performed comparisons across aggregations of treatment groups corresponding to sixty effects of interest defined in our primary analysis and found suitable balance for each aggregation.

[71] For example, one drawback of the linear model framework is the requirement that the causal estimands be defined as parameters of the probability distribution of the observed response. To the contrary, our results do not rely on distributional assumptions. *See 2^K Factorial Designs*, *supra* note 54.

[72] *See* Jerzy Splawa-Neyman, *Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych* [*On the Application of Probability Theory to Agricultural Experiments*], 10 ROCZNIKI NAUK ROLNICZYCH 1 (1923), *translated in* 5 STAT. SCI. 465 (1990).

[73] *See* Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS'N 945 (1986).

[74] Under the RCM, each unit (i.e., each participant) in this experiment has twenty-two "potential outcomes," one for each possible treatment combination. For example, a participant may have awarded $4 million had he been randomized to the punitive damages scenario of the control condition, $2 million had he been randomized to the unbiased, low-variability, average, punitive damages treatment condition, and so on and so forth for all twenty-two possible treatment combinations. The RCM frames causal inference as a missing-data problem: because we can observe only

Using this framework, we defined "estimands" (quantities of interest) for effects on accuracy, spread, and magnitude as well as corresponding "estimators" (functions of the data that we used to estimate the quantities of interest).[75] We also applied secondary definitions to determine whether an observed effect is robust to other reasonable definitions and to obtain a more complete picture of the effects of interest.

one potential outcome for each unit—the one to which the unit was in fact assigned—we do not know, and therefore must estimate, the values of the unobserved potential outcomes to make causal conclusions. *See generally 2^K Factorial Designs*, *supra* note 54; Holland, *supra* note 73; Rubin, *supra* note 54.

[75] We sought to assess the finite-population effects of prior-award information for different levels of bias, variability, and form on the accuracy, spread, and magnitude of resulting award values, where accuracy is defined as the proximity of the awards to the "correct" award in terms of both bias and variance. These objectives motivated our choice of estimands. We used the mean of the logarithm-transformed data to define magnitude, the interquartile range (IQR) (i.e., the difference between the 75th percentile and the 25th percentile), to define spread, and the mean squared error (equal to the sum of the variance and the squared bias relative to the "correct" award) to define error. An estimand for the effect of a certain variable on spread, for example, can therefore be defined as the difference between the IQR (in terms of potential outcomes) associated with one comparison group and that of another comparison group—for example, the IQR associated with the unbiased punitive damages group and that of the biased punitive damages group. A corresponding estimator can then be defined using the observed data. Using the factorial-effect estimators we estimated an effect by computing the values of accuracy, spread, and magnitude for each of the treatment combinations involving the factor (or combination of factors) of interest, and then separately averaging over each of the treatment combinations with the same level of the factor (or combination of factors). For example, for the main effect of bias on magnitude within the pain and suffering scenario, we: 1) calculated the magnitude of award amounts separately for each of the ten treatment combinations receiving either biased or unbiased prior award information in the pain and suffering scenario using the mean of the logarithm-transformed values; 2) averaged the values from step (1) across the five treatment combinations receiving biased information, and separately averaged the values from step (1) across the five treatment combinations receiving unbiased information; and 3) calculated the difference of the averages obtained in step (2). Also note that the estimands and estimators defined in this subsection applied to main effects as well as all two-way, three-way, and four-way interaction effects.

We were interested in testing the null hypothesis that the true difference between comparison groups is zero. Once we calculated the estimates of factorial effects using observed data, we sought to understand the "statistical significance" of the estimated effects, which can be interpreted as an indicator of how unlikely the observed differences between treatment groups would be under certain hypothesized effects.[76] To do this, we used the Fisherian approach[77] and applied randomization tests to evaluate "Fisher's sharp null hypothesis" of "no unit-level treatment effects." We applied this approach to approximate "Fisher exact p-values" for each hypothesis and to generate 95% "Fisher intervals" for certain effects of interest.[78] All *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure to control the

[76] Rubin, *supra* note 54.

[77] Note that the RCM allows such inference using the Neymanian perspective, Fisherian perspective, or the Bayesian perspective. *See* Donald B. Rubin, *Bayesian Inference for Causal Effects: The Role of Randomization*, 6 ANNALS STAT. 34 (1978). Because the estimator for capturing, for example, the spread of award values is not an unbiased estimator of its corresponding estimand, and for other reasons, we decided not to use Neyman's method. One of the major advantages of the Fisherian approach to inference, as opposed to a model-based or Bayesian approach, is that it does not require any assumptions about the underlying distribution of the data. Thus, randomization tests can be constructed and applied for any test statistic, regardless whether its distribution is known.

[78] Specifically, for each effect of interest, we assumed the "sharp null" hypothesis of "no unit-level treatment effect" to impute missing potential outcomes for each participant. We then generated a sample of N_{sim}=250,000 possible randomizations of the *N* participants. *See generally* GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES (2015). For each of these 250,000 possible randomizations, we recomputed the value of the test statistic under the (hypothetical) randomization. The resulting distribution is referred to as the "randomization distribution" of the test statistic. We approximated Fisher exact *p*-values for each hypothesis (to compute statistical significance) by calculating the proportion of values in the randomization distribution that were equal to or more extreme than the observed test statistic. For certain effects, we generated 95% Fisher intervals by calculating a sequence of (unadjusted) *p*-values corresponding to a null hypothesis other than the hypothesis of "zero treatment effect," and then identified the hypothesized values that result in *p*-values equal to or larger than 0.05. For all estimates, we assumed a constant additive treatment effect.

FDR.[79] After adjustment, we considered *p*-values of less than α=0.05 to be statistically significant. "Significance" suggests that there is substantial evidence within the data against the null hypothesis of no treatment effect.[80]

E. Leg II Experimental Design

The primary purpose of Leg II was to test certain questions that arose from Leg I. First, Leg I tested the effects of prior-award information that involved a downward bias, but not prior-award information that involved an upward bias—where the median of the prior awards was substantially above the median of the control group. Second, certain treatment arms in Leg I involved prior-award information that was heavily right-skewed in distribution, which may have introduced certain effects. Therefore, in Leg II, we aimed to test 1) the effects of unbiased prior-award information when prior awards were approximately symmetric rather than right-skewed; and 2) the effects of upwardly biased prior-award information (with minimal or no skew). We also sought to test the effects of prior-award information using alternative definitions of accuracy, spread, and magnitude.

We generally applied the design of Leg I to Leg II,[81] only altering the numerical values provided to participants. Also, Leg II, which used only the list form for prior-award information, involved only ten treatment conditions and a significantly reduced sample size. As in Leg I, we determined the numerical values of prior awards using a pilot study that substantively replicated the control groups. In Leg II, we constructed approximately symmetric distributions around the medians observed in

[79] *See* Benjamini & Hochberg, *supra* note 70; Maria T. Kimel et al., *The False Discovery Rate for Multiple Testing in Factorial Experiments*, 50 TECHNOMETRICS 32 (2012). Due to having a large number of tests, we divided our tests into primary analyses and secondary analyses, and we applied adjustments for multiple comparisons within each category and within each level of scenario (which we treated as separate studies for purposes of our analysis).

[80] Note that, throughout the study, we made the "stable unit treatment value assumption" (SUTVA). This means 1) that the potential outcome of a unit depends only on its own assignment, and not on the assignments of other units, and 2) that there are no "hidden versions" of treatment. *2^K Factorial Designs*, *supra* note 54, at 730. We believe this was justified.

[81] This includes procedures for assessing covariate balance, which our randomization achieved.

the pilot study ($500,000 for punitive damages and $2 million for pain and suffering); and we multiplied the unbiased prior-award values by approximately three to arrive at biased prior-award values. Table B provides a list of the Leg II treatment conditions and corresponding prior-award information.

Table B. Prior-Award Information Provided in Leg II Treatment Conditions

Condition	Prior-Award Information
1. Pain and suffering control	[None]
2. Pain and suffering LV unbiased	1.2m, 1.7m, 2m, 2.4m, 2.9m
3. Pain and suffering HV unbiased	120k, 1.2m, 2m, 2.8m, 3.9m
4. Pain and suffering LV biased	3.6m, 5.1m, 6m, 7.2m, 8.7m
5. Pain and suffering HV biased	360k, 3.6m, 6m, 8.4m, 11.7m
6. Punitive damages control	[None]
7. Punitive damages LV unbiased	350k, 410k, 500k, 590k, 680k
8. Punitive damages HV unbiased	10k, 210k, 500k, 720k, 1.1m
9. Punitive damages LV biased	1.2m, 1.3m, 1.5m, 1.7m, 1.9m
10. Punitive damages HV biased	30k, 630k, 1.5m, 2.2m, 3.2m

Importantly, in addition to the purposes stated above, Leg II was intended to test certain limits of prior-award information in improving the accuracy of damage awards. We used, for example, relatively extreme values in the biased conditions and smaller differences between the high-variability and low-variability conditions. We also used a substantially smaller sample size. Therefore, as we noted up front in our Leg II design, any failure to observe the tested effects in Leg II does not negate or call into question our observed effects in Leg I.

IV. RESULTS

Our analysis involved approximately 400 hypothesis tests in Leg I and 250 hypothesis tests in Leg II, with numerous tests informing a particular effect.[82] Our general approach to interpreting the data was as follows: First, we examined the comparison that most directly informs an effect of interest. Second, we examined other comparisons that inform the effect less directly. Third, if we found support for the effect in both steps (1) and

[82] Note that we followed strict procedures to ensure that we could not access outcome data prior to completing the design phase of each leg of the experiment, respectively.

(2), we generally interpreted the data as strongly evidencing an effect. If we found support for the effect in step (1), but an absence of evidence for the effect in step (2), we generally interpreted the data as evidencing an effect, unless there was a particular reason to believe otherwise. If we found support for the effect in step (1) but evidence *against* it in step (2), we examined the inconsistency that arose in step (2). If our examination revealed a straightforward explanation that defused the inconsistency, we interpreted the data as evidencing an effect and sought to confirm the effect in Leg II. If the examination failed to reveal a straightforward explanation, we interpreted the data as not evidencing an effect, and we examined the results for an alternative explanation.

A. The Data

There are approximately 240-250 observations for each treatment level.[83] To address extreme outliers, we "winsorized" the data at the 99th percentile for our primary analysis and analyzed the effects of prior-award information on outliers—defined as awards above the 99th percentile in each scenario—separately. This means that any awards above the 99th percentile—$50 million for punitive damages and $100 million for pain and suffering—were recoded to the award value of the 99th percentile.[84] We took this step because extreme outliers can cause misleading results

[83] Prior to analyzing the data, we applied an initial "data cleaning" procedure to ensure data quality. First, to ensure the validity of responses (e.g., that they originated from registered Mechanical Turk users who met our inclusion criteria), we excluded twenty-three participants who entered incorrect payment codes. Second, we excluded four participants who provided award amounts that were deemed nonsensical by our software. Because only a very small number of participants were excluded from the initial sample of 5,500 participants, and because there was no evidence suggesting any resulting systematic distortion, we excluded these participants without applying advanced missing data techniques. Additionally, as indicated in our methodology discussion, *supra* Part III, participants who marked "Other" for their sex were excluded due to our inability to sensibly merge that category with one of the other categories of sex for purposes of testing for covariate balance. After applying these exclusion criteria, our final Leg I sample size is N=5,458.

[84] Similarly, in Leg II, we winsorized the data at the 99th percentile ($50 million and $174.6 million for punitive damages and pain and suffering, respectively).

that are dominated by chance rather than true effects.[85] Furthermore, to ensure the robustness of our findings, we also tested them under alternative winsorization thresholds. We report the corresponding results below.

Thus, unless stated otherwise, the results we report below reflect datasets winsorized at the 99th percentile. In certain instances, we report results from alternative winsorization schemes for illustrative purposes or to make a point regarding those data in particular, but we identify such instances clearly. Summary statistics for data in Leg I and II are provided in Tables C and D, respectively.

[85] Winsorizing had minimal consequences for our results in the punitive damages scenario (since outliers in that scenario are less extreme in size and effect) but was important to prevent spurious results in the pain and suffering scenario (although ultimately it was not dramatically consequential for our conclusions). Note that our original design in Leg I did not include a winsorization scheme due to our sensitivity to the importance of outliers in this study, but due to a few extreme outliers, it became clear that not using such a method would lead to spurious results. We therefore decided to use the conservative method described in the text and examine the effect of prior-award information on outliers separately. Note that such extreme awards (and probably some that are far less extreme) are unlikely in practice, and would be subject to reduction by the courts using the device of remittitur.

Table C. Summary Statistics for Leg I Sample (*N* = 5,458)

Statistic	Punitive Damages (Raw)	Punitive Damages (Winsorized)	Pain and Suffering (Raw[86])	Pain and Suffering (Winsorized)
Sample Size	2751	2751	2704	2707
Mean (millions)	$18.3	$2.3	$23.8	$6.7
Median (millions)	$0.5	$0.5	$3.0	$3.0
SD (millions)	$397.3	$6.4	$396.0	$14.0
IQR (millions)	$0.9	$0.9	$3.5	$3.5

Table D. Summary Statistics for Leg II Sample (*N* = 2,521)

Statistic	Punitive Damages (Raw)	Punitive Damages (Winsorized)	Pain and Suffering (Raw)	Pain and Suffering (Winsorized)
Sample Size	1262	1262	1259	1259
Mean (millions)	$13.1	$2.35	$23.8	$8.5
Median (millions)	$1.0	$1.0	$4.0	$4.0
SD (millions)	$286.6	$6.2	$255.2	$20.3
IQR (millions)	$1.4	$1.4	$6.4	$6.4

B. Effect on Accuracy

The data provide strong evidence that prior-award information reduces error and improves accuracy. In both levels of scenario*,* and across all levels of bias and variability, and their interactions, prior-award information had a significant negative effect on error and significant positive effect on accuracy, which we measured using mean squared error

[86] For descriptive purposes, we exclude from this summary of raw pain and suffering data three extreme values that are greater than or equal to $100 billion.

(MSE), a combination of variance and bias. Figure 1 illustrates the approximate randomization distributions and observed test statistics for the hypothesis tests of "no treatment effect" in each scenario; Figures 2 and 3 summarize the effects on error (the inverse of accuracy) with 95% Fisher intervals and levels of significance. Note that, throughout this Part, we report unadjusted *p*-values and use stars to indicate statistical significance *after* correction for multiple comparisons.[87]

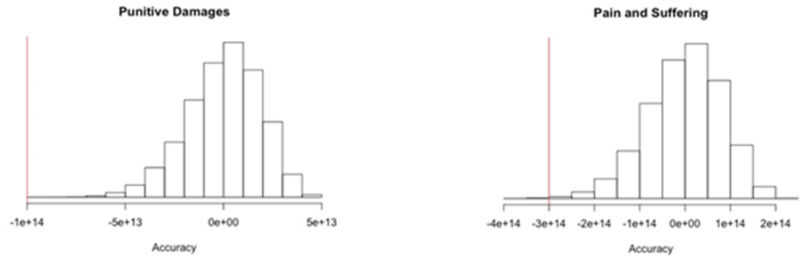

Figure 1- Randomization Distributions. Randomization distributions for effect of treatment (any prior-award information) versus control (no prior-award information) on accuracy for punitive damages (left) and pain and suffering (right). Red lines show observed test statistics.

[87] *** denotes statistical significance at the α=0.001 level, ** denotes statistical significance at the α=0.01 level, and * denotes statistical significance at the α=0.05 level.

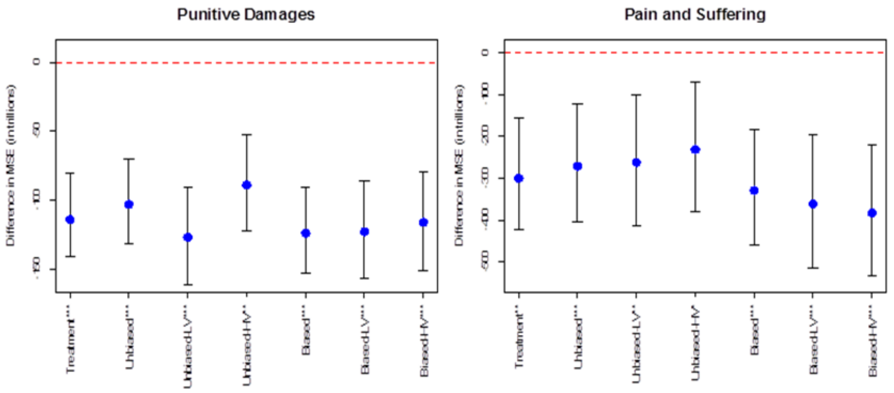

Figure 2 – Effects on Error. Point and interval estimates for difference in mean squared error (MSE) (in trillions) for each treatment condition for punitive damages (left) and pain and suffering (right) vs. control ($146 and $492, respectively). Stars indicate statistical significance (after correction for multiple comparisons) of the difference in MSE between each treatment condition and control.

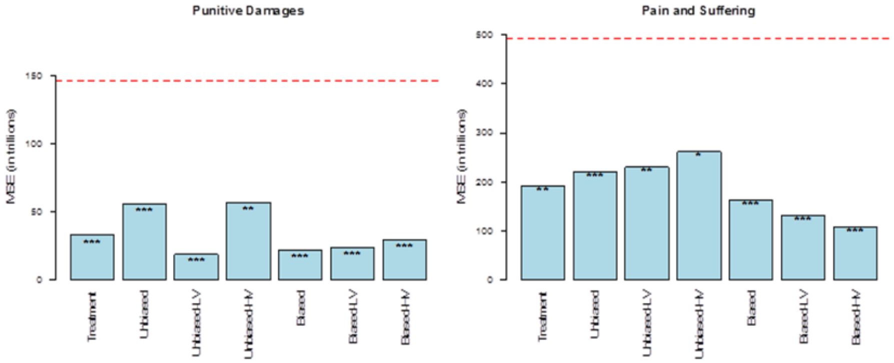

Figure 3 – Effects on Error. Observed mean squared error (MSE) (in trillions) for different treatment combinations for punitive damages (left) and pain and suffering (right). Dotted lines at $146 and $492 show MSE for control groups, respectively, and stars indicate statistical significance (after correction for multiple comparisons) of difference in MSE between each treatment condition and control.

As discussed above, prior-award information may introduce error in the form of distortion, or bias, but also reduce error by reducing the dispersion of awards. We discuss effects on spread and magnitude below.

The results above show, however, that any distortionary effects on the size of awards are dominated (in terms of error) by the beneficial effects on dispersion. In particular, prior-award information caused an improvement in accuracy across all levels of bias, variability, and form—effects largely observed at the .001 level (and all but one at the .01 level). Furthermore, we observed similar results under secondary winsorization schemes and definitions of accuracy. For example, to confirm that our results held without the influence of more-extreme values, we examined four important effects using data winsorized at the 90th percentile—i.e., $5 million for punitive damages and $10 million for pain and suffering. Using these data, we again observed significant positive effects on accuracy, all at the .001 level. Similarly, we observed positive effects on accuracy using mean *absolute* error (MAE) rather than MSE to define accuracy*.* Our results for these effects are summarized in Figures 4-6.

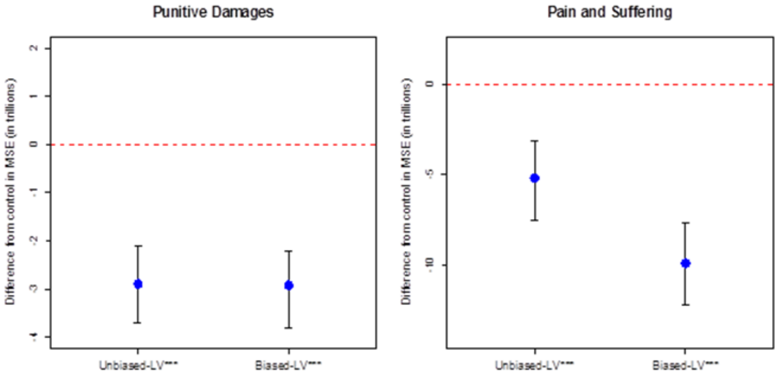

Figure 4 - Effects on Error When Winsorizing at the 90th Percentile. Point estimates and 95% Fisher intervals for difference in mean squared error (MSE) (in trillions) vs. control for unbiased low-variability and -biased low-variability conditions for punitive damages (left) and pain and suffering (right) when winsorizing at the 90th percentile in each scenario.

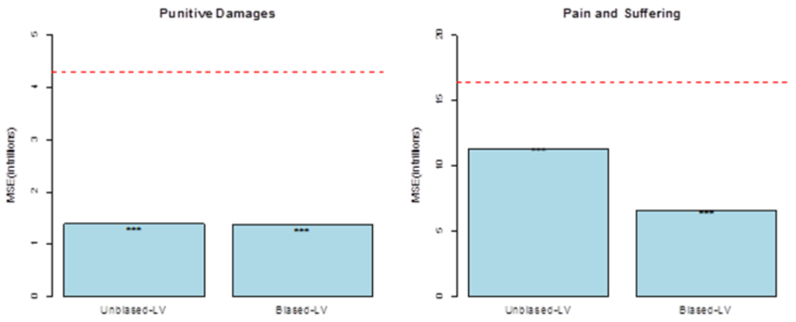

Figure 5 - Effects on Error When Winsorizing at the 90th Percentile. Observed mean squared error (MSE) (in trillions) for unbiased low-variability and biased low-variability conditions for punitive damages (left) and pain and suffering (right) when winsorizing at the 90th percentile in each scenario. Dotted lines show MSE for control groups, and stars indicate statistical significance (after correction for multiple comparisons) of difference in MSE between each treatment combination and control.

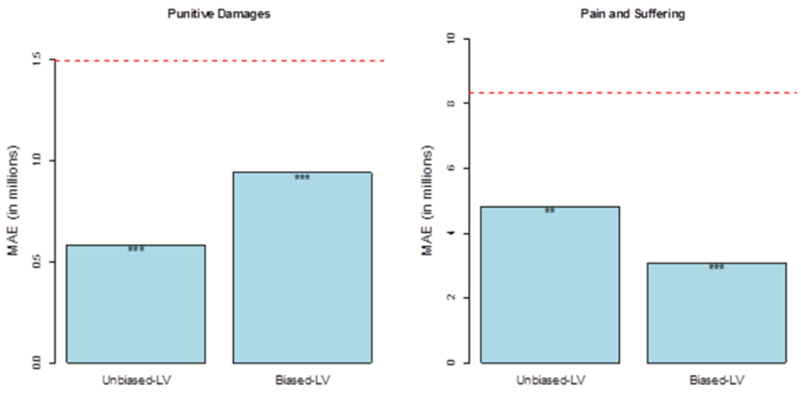

Figure 6 - Effects on Error When Using MAE and Winsorizing at the 90th Percentile. Observed mean absolute error (MAE) (in trillions) for unbiased low-variability and biased low-variability conditions for punitive damages (left) and pain and suffering (right) when winsorizing at the 90th percentile in each scenario. Dotted lines show MAE for control groups, and stars indicate statistical significance (after correction for multiple comparisons) of difference in MAE between each treatment combination and control.

We observed similar results in Leg II, with positive effects on accuracy throughout all levels of scenario, bias, and variability, aside from two tests that indicated positive but non-significant effects on accuracy (see Figure 7). Additionally, aside from two benign exceptions, all tests using a 90th percentile winsorization scheme or MAE for the definition of accuracy resulted in significant positive effects on accuracy.[88]

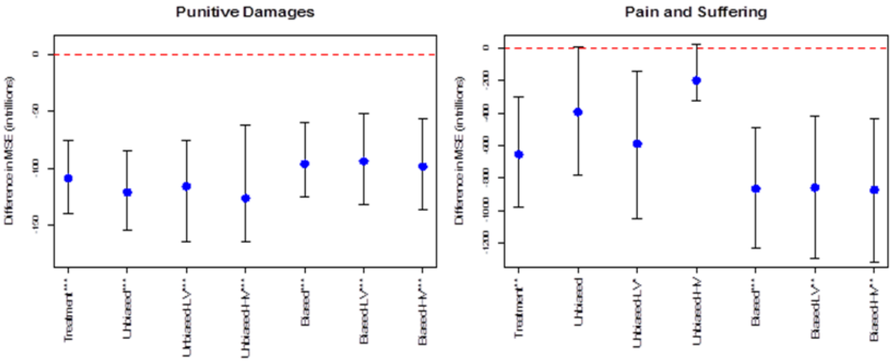

Figure 7 – Leg II Effects on Error. Point and interval estimates for difference in mean squared error (MSE) (in trillions) vs. control for each treatment condition for punitive damages (left) and pain and suffering (right) in Leg II.

[88] Throughout this Part, aside from a few comments, we do not focus on the effects of form on the various outcome variables. The most important finding with respect to form is that our results are generally not sensitive to the particular form of the prior-award information, and specifically, to whether we provided jurors with prior-award information in the form of an average, range, or list. Most significantly, we found that prior-award information improves accuracy regardless of the level of form. More broadly, we found that form generally did not have a significant effect on the outcome variables studied. This does not imply that prior-award form necessarily would have no significant impact on jury determinations. *See generally supra* note 53; Saks et al., *supra* note 22. Rather, we simply did not detect a significant impact in our study. It is possible, for example, that such effects were rendered undetectable in our study by other sources of variability. Nevertheless, our findings with respect to form arguably suggest that if the development of one form of prior-award information involves lower costs than others, it could be cost-effective to present prior-award information to a jury in this form, in light of the similar levels of effectiveness detected across various forms. However, further research regarding prior-award form would be valuable, and necessary to draw a firm conclusion in this regard.

C. Effect on Spread

The effects of prior-award information on accuracy provide evidence of the effects on the dispersion of awards as well. Specifically, because accuracy is defined using MSE, which can be deconstructed into bias and variance, and because prior-award information can only add bias, and not reduce it, we know that any improvement in accuracy is due to a reduction in variance. However, we separately tested the effect of prior-award information on spread, which we defined using the IQR rather than variance. Understanding the effect of prior-award information on spread, in addition to its effect on accuracy, permits a more nuanced understanding of the data.

The IQR is, in a sense, less sensitive to differences in random variation. For example, changing a value at the 95th percentile from $1 million to $50 million would not affect spread. But spread provides specific information—the difference between the 75th and 25th percentiles—that may be obscured in other measures of dispersion.

The data provide strong evidence that prior-award information reduces spread. In the punitive damages scenario, our comparison of all active treatment conditions (combined) to the control condition indicated that, overall, prior-award information caused a significant reduction in spread (p<0.001***). Furthermore, we found that both unbiased prior-award information and biased prior-award information caused a reduction in spread (p=0.02* for unbiased and p<0.001*** for biased).[89]

In several tests, particularly in the pain and suffering scenario, we detected a significant increase in spread. This result was not very surprising, however, because the range of the prior awards—and particularly the high-variability prior awards—was frequently substantially larger than the control group IQR.[90] This caused spread to

[89] Note that the average level of form tended to have a greater downward impact on spread relative to the other levels of form. This may be interpreted as resulting from the participants' perception that there is no variability in the prior-awards, since they were provided only a single number. Alternatively, the participants could have interpreted the average form as providing less information and therefore "deserving of" less influence. Participants simply did not know whether the average reflected, for example, five prior awards of identical values or five highly scattered prior awards. On balance, they seemed to have "interpreted" (implicitly) the information as reflecting awards of lower variability.

[90] There are two reasons for this: first, the percentiles chosen for determining the range of values of unbiased high-variability prior awards are the 15th

remain unchanged (i.e., without significant effect), and even to increase in response to high-variability prior-award information, notwithstanding an overall reduction in dispersion, as reflected in the effects of unbiased prior-award information on accuracy.[91] Effects on spread, with Fisher intervals and levels of significance, are summarized in Figure 8.

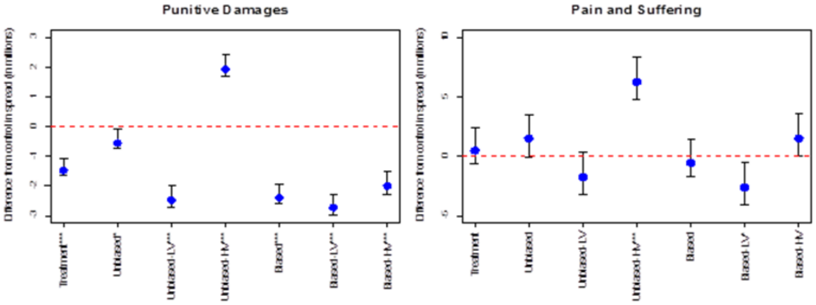

Figure 8 – Effects on Spread. Point estimates and 95% Fisher intervals for difference in spread (in millions) vs. control for each treatment condition for punitive damages (left) and pain and suffering (right).

and 85th percentiles, substantially wider than the IQR's 25th and 75th percentiles; second, there was greater dispersion in the pilot study data than in the control group data, caused by "choppiness" in the data or sampling variation. In Leg II, we addressed this issue by testing the effect on spread using narrower prior-award distributions. *See infra* Figure 9.

[91] For example, unbiased high-variability and unbiased low-variability prior awards in the pain and suffering scenario ranged from $200,000 to $15 million and from $1.12 million to $4.7 million, respectively, compared to the 25th and 75th percentiles of the pain and suffering control group awards, which were $50,000 and $5 million, respectively. This explanation is corroborated by our results in the punitive damages scenario, where, although spread decreased in response to unbiased and biased prior-award information (separately and combined), and decreased in response to unbiased low-variability, biased low-variability, and biased high-variability prior-award information, it increased in response to unbiased high-variability prior-award information. This makes sense, because, as with the pain and suffering scenario, unbiased high-variability prior awards ranged from $10,000 to $10 million, a range far greater than the control group 25th and 75th percentiles ($100,000 and $3 million, respectively). Compare this to the unbiased low-variability prior awards, which ranged only from $500,000 to $1 million, well within the range of control group 25th and 75th percentiles. This interpretation is confirmed by our results in Leg II, which used narrower prior-award distributions and resulted in significant reductions in spread across the board.

To confirm our interpretation, we first tested effects on spread using alternative winsorization thresholds and measures of dispersion. Specifically, we tested the effects of unbiased prior-award information on the standard deviation of awards using data winsorized at the 90th percentile. These tests provide substantial support for our interpretation above. Specifically, in both levels of scenario, unbiased prior-award information reduced the standard deviation of awards (p<0.001*** for punitive damages and for pain and suffering). Further, in both levels of scenario, unbiased *low-variability* prior-award information reduced the standard deviation of awards (p<0.001*** for punitive damages and for pain and suffering), whereas unbiased *high-variability* prior award information had no significant effect due to the relatively high dispersion of unbiased high-variability prior awards.

Moreover, we tested and confirmed our interpretation using Leg II. In Leg II, which involved substantially narrower distributions of prior awards, across all levels of scenario, variability, and bias, prior-award information reduced spread (almost always at the .001 level) using our standard definition of spread (IQR) and our standard winsorization threshold (99th percentile).[92] The Leg II effects on spread are illustrated in Figure 9.

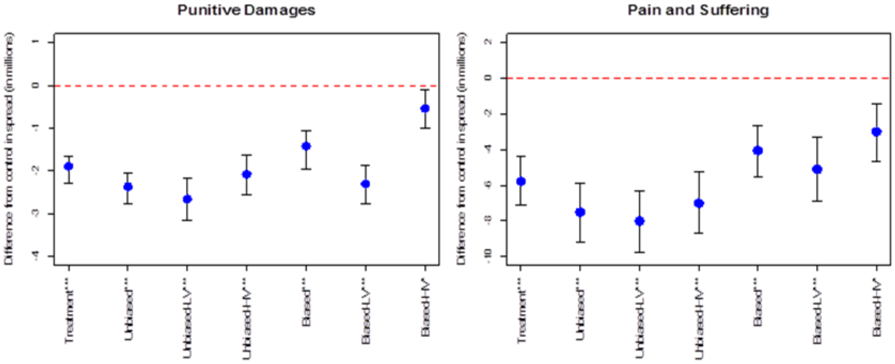

Figure 9 – Leg II Effects on Spread. Point estimates and 95% Fisher intervals for difference in spread (in millions) vs. control for each

[92] In addition, we found these effects to be robust to alternative definitions and winsorization schemes: aside from a very small number of tests resulting in negative but non-significant effects, defining spread using variance and winsorizing at the 99th percentile or the 90th percentile resulted in significant reductions in spread across all levels of scenario, variability, and bias.

treatment condition for punitive damages (left) and pain and suffering (right) in Leg II.

D. Effect on Magnitude

The effect of prior-award information on magnitude is less straightforward than its effects on accuracy and spread. Using our standard definition of magnitude, the mean of the log-transformed data, we found that unbiased prior-award information generally caused a positive effect on magnitude for all levels of scenario, variability, and form. However, using the median to define magnitude, we found no effect, and using the mean to define magnitude, we found no effect or a negative effect.

We can explain these disparities by considering the differences among the various measures. The mean of the log-transformed data is a commonly used measure for testing magnitude with right-skewed data. The reason that it is popular for right-skewed data is that it "pulls in" extreme values more than it "pulls in" moderate values. On the other hand, the median, which accounts only for differences in order, is not at all affected by the skew; and the mean, which is sensitive to size, is heavily influenced by extreme values. We can therefore understand the data as follows: unbiased prior-award information had no significant effect on the size of awards, in the sense that the central award—the median—was not affected. For example, in the punitive damages scenario, the median of awards in the control condition is equal to the median of awards in unbiased conditions, $1 million. This result supports the argument that providing jurors with unbiased prior-award information reduces the dispersion of awards while distorting award size (measured using the median) minimally or not at all. Furthermore, in line with our expectations, unbiased prior-award information had no effect or a negative effect on award size when measured using the mean, thus reflecting more extreme but non-outlier awards (when winsorizing at the 99th percentile) or reflecting all outlier and non-outlier awards (when not winsorizing).[93] This interpretation was corroborated by our exploratory

[93] Zero and negative effects are in line with our expectations when defining magnitude using the mean. Because the distribution of award determinations is heavily right-skewed (whether winsorizing at the 99th percentile or not), we would expect that a reduction in dispersion (including, for example, a reduction in extreme values) would also have a negative impact on the mean award.

analysis regarding outliers.[94] Finally, however, unbiased prior-award information had a positive effect on award size when measured using the mean of the log-transformed data, thus reflecting the effects—but greatly diminished effects (through the log transformation)—of more extreme but non-outlier awards. This positive effect results, in a sense, from weighting non-extreme values more heavily than more extreme (but non-outlier) values.

We confirmed our findings and interpretation using Leg II of the experiment, where we observed similar effects—that is, positive effects or no effects on magnitude using the mean of the log-transformed data and no effects or negative effects using the median or mean of the (unlogged) data.[95] We summarize the effects of unbiased prior-award information on magnitude (based on winsorization at the 99th percentile unless stated otherwise) using various definitions in Table E below.

Table E. Effect of Unbiased Prior-Award Information on *Magnitude* Under Alternative Definitions[96]

	Punitive Damages				Pain and Suffering			
	Log mean	Median	Mean	Mean (winsor 90th)	Log mean	Median	Mean	Mean (winsor 90th)
Unbiased vs control	Positive	0	Negative	0	Positive	0	0	Positive
Leg II: Unbiased vs control	0	0	Negative	Negative	Positive	0	Negative	Negative

Based on our results in Table E, the effect of unbiased prior-award information on magnitude is highly sensitive to the measure used to define it. In any event, although the data evidence no effect using various common measures of central tendency, such as the median, the observed effects on magnitude using the mean of the log-transformed data may have

[94] *See infra* Section IV.F.

[95] Note that the prior awards in Leg II did not involve a substantial right skew, which may have exacerbated any positive effects on magnitude observed in Leg I.

[96] For comparative purposes, Leg II results in Table E are generally listed for winsorization at the 99th percentile. It is important to note, however, that using the mean of the log-transformed data and winsorizing at the 95th percentile (a standard definition for magnitude in Leg II) results in positive effects for punitive damages and pain and suffering.

implications. For example, it is not unlikely that the upper bound of a list of prior awards has a positive anchoring effect on award determinations.[97]

Finally, in line with our expectations, biased prior-award information impacted the magnitude of awards, relative to the magnitude of awards in unbiased conditions, in the direction of the bias. That is, for all levels of scenario and variability, and in both Leg I and Leg II, downward bias had a negative effect on magnitude and upward bias had a positive effect on magnitude, relative to the magnitude observed in unbiased conditions.

E. Effect of Variability

We observed strong evidence that variability has significant influence on the impact of prior-award information. The data indicate that low-variability prior-award information had a greater impact on spread and magnitude—whether the impact was negative or positive—than did high-variability prior-award information.

Pursuant to the model in *The Logic of CCG*, low-variability prior-award information provides *more* information than high-variability prior-award information[98] and thus has greater influence on award determinations. This "influence" translates to greater impacts on spread and magnitude. Consistent with this model, for unbiased prior-award information, variability had no significant effect on magnitude (since the prior-award information was unbiased) but low-variability prior-award information had a significantly greater impact (negative impact) on spread than did high-variability prior-award information. That is, relative to unbiased high-variability prior-award information, unbiased low-variability prior-award information had no significant effect on magnitude and a significant negative effect on spread (p<0.001*** for punitive damages and for pain and suffering). Moreover, biased low-variability prior-award information had a significantly greater impact on both magnitude and spread than did biased high-variability prior-award information. In other words, relative to biased high-variability prior-award information, biased low-variability prior-award information had a significant negative effect on both magnitude (p<0.001*** for punitive damages and for pain and suffering) and spread (p=0.004** for punitive damages and p<0.001*** for pain and suffering). Our results regarding the effects of low-variability prior-award information relative to high-variability prior-award information are summarized in Table F below.

[97] We discuss implications in Part VI.

[98] *See supra* Part II.

Table F. Low-Variability vs. High-Variability Prior-Award Information

		Punitive Damages		Pain and Suffering	
		Observed test statistic	p-value	Observed test statistic	p-value
Unbiased: low variability vs high variability	Spread	-$4,397,500	0.000***	-$8,000,000	0.000***
	Magnitude	-0.09	0.591	-0.18	0.165
Biased: low variability vs high variability	Spread	-$727,500	0.004**	-$4,112,500	0.000***
	Magnitude	-0.67	0.000***	-0.52	0.000***

Note that a possible argument against this interpretation is that the results observed can be explained by the effect of high variability on spread rather than the effect of low variability on the influence of the prior-award information. Specifically, as suggested above, it is possible that high-variability prior-award information had a positive effect on spread, since the range of the high-variability prior awards was greater than the IQR of awards. Arguably, therefore, the observed effect here results only from removing the upward effect on spread caused by this high-variability prior-award information. There are, however, two counterarguments. First, this argument cannot explain the effects of biased low-variability prior-award information relative to biased high-variability prior-award information, since neither of these conditions had a positive effect on spread relative to the control condition. Furthermore, remember that biased low-variability prior-award information did not involve prior awards that were of lower amount than biased high-variability prior-award information; rather, it had only lower variability. The biased low-variability prior-award information (ranging from $50,000 to $100,000) was completely contained within the range of values in the biased high-variability prior-award information (ranging from $1,000 to $1 million). Thus, the significant negative effects of biased low-variability relative to biased high-variability prior-award information provide particularly strong evidence that low-variability prior-award information had greater influence—i.e., was more impactful—on award determinations than did high-variability prior-award information.

Second, we confirmed our interpretation using Leg II. Specifically, Leg II involved prior-award distributions that were narrow and approximately symmetric; and, although we did not have sufficient power (as discussed above, due to relatively small differences among prior-award distributions and relatively small sample sizes)[99] to detect many effects with respect to variability, the data provide strong evidence in the punitive damages scenario that low-variability prior-award information indeed had a significantly greater negative impact on spread than did high-variability prior-award information.

F. Effect on Outliers

In our exploratory analysis, we examined the effect of prior-award information on "outliers," which we defined as awards at or above the 99th percentile ($50 million in the punitive damages scenario and $100 million in the pain and suffering scenario). We found strong evidence that prior-award information had a significant negative effect on outliers.

In the control group of the punitive damages scenario, there were thirteen observations out of 249 (5.2%) above $50 million, whereas in the ten treatment arms of the punitive damages scenario, there were twenty-three observations out of 2502 (0.9%) above $50 million, or, on average, approximately two observations (23/10 = 2.3) above $50 million per 249 observations. This represents a significant reduction in outliers (*p*<0.001**)*.* Furthermore, in the five unbiased treatment arms in the punitive damages scenario, there were fourteen observations out of 1257 (1.1%) above $50 million, or, on average, approximately three (14/5 = 2.8) above $50 million per 249 observations. This also represents a significant reduction in outliers (*p*<0.001***).

In the control group of the dataset for the pain and suffering scenario, there were ten observations out of 244 (4.1%) above $100 million, whereas in the ten treatment arms of the pain and suffering scenario, there were thirty-three observations out of 2463 (1.3%) above $100 million, or, on average, approximately three observations (33/10 = 3.3) above $100 million per 249 observations. This represents a significant reduction in outliers (*p*=0.002**). Furthermore, in the five unbiased treatment arms in the pain and suffering scenario*,* there were nineteen observations out of 1229 (1.6%) above $100 million, or, on average, approximately four (19/5 = 3.8) above $100 million per 249 observations. This also represents a significant reduction in outliers (*p*=0.018*)*.* Our analysis regarding the

[99]*.* *See supra* Section III.E.

effect of prior-award information on outliers is summarized in Table G and Figure 10.[100]

Table G. Effect on Outliers

	Punitive Damages (Raw)			Pain and Suffering (Raw)		
	Treatment rate	Control rate	p-value	Treatment rate	Control rate	p-value
Treatment vs Control	0.9%	5.2%	0.000***	1.3%	4.1%	0.002**
Unbiased vs Control	1.1%	5.2%	0.000***	1.6%	4.1%	0.018*

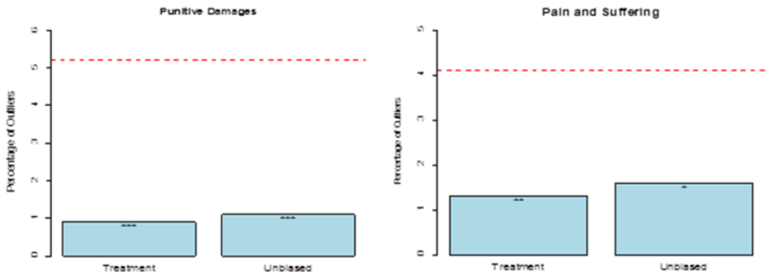

Figure 10 – Effect on Outliers. Percentage of outlier awards for treatment and unbiased conditions for punitive damages (left) and pain and suffering (right). Dotted lines show percentage of outliers for control, and stars indicate statistical significance (after correction for multiple comparisons) of the difference in percentage of outliers between each treatment combination and control.

[100] Our Leg II results confirmed our findings in Leg I. In Leg II, we observed significant negative effects on outliers for all tests in the punitive damages scenario (treatment vs. control, unbiased vs. control, and biased vs. control). For the pain and suffering scenario, we observed significant negative effects on outliers caused by biased prior-award information, and reductions—although non-significant reductions—in outliers associated with unbiased prior-award information and overall prior-award information. Note that these latter two non-significant results are due to unusually low power and do not affect our other results regarding the impact of prior-award information on outliers. For example, were we to define an "outlier" as the top 2% of awards (which would allow for a larger sample of "outliers"), these tests would be significant as well (p<.001*** for treatment vs. control and p=0.029* for unbiased vs. control).

G. Effect on Explanations

In our exploratory analysis, we examined the effect of prior-award information on explanation response rates. As described in Part III, we provided participants with the option of providing an explanation for the amount they awarded. In total, 78.8% of participants who received the punitive damages scenario and 76.2% of participants who received the pain and suffering scenario provided explanations for their awards.

We found strong evidence that prior-award information had a significant positive effect on explanation response rates. Specifically, 78% of participants who received prior-award information provided explanations, whereas only 72% of participants who did not receive prior-award information—those who were assigned to control groups—provided explanations, representing a significant positive effect on the overall explanation response rate (*p*=0.0027****). Furthermore, we observed a significant positive effect on the response rate in the pain and suffering scenario (*p*=0.004*) and a positive but non-significant effect on the response rate in the punitive damages scenario (*p*=0.083).[101]

V. Limitations

Before discussing the implications of our results, we highlight a number of important limitations. First, the experimental units in this study were mock *jurors* rather than mock *juries*. Although juries are composed of jurors, there is concern that deliberation among jurors would cause jury awards to differ from juror awards, thereby limiting the applicability of juror behavior to draw conclusions regarding jury behavior.

Numerous studies have shown that predeliberation juror preferences or certain aggregations of juror preferences serve as good predictors of jury awards.[102] Nevertheless, there is a justified concern that the random

[101] In Leg II, prior-award information was associated with higher explanation response rates, but due to substantially smaller sample sizes (and, in some instances, smaller increases), these effects were not significant at the 0.05 level.

[102]. *See* DENNIS J. DEVINE, JURY DECISION MAKING: THE STATE OF THE SCIENCE 176-77 (2012); Shari Seidman 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); s*ee also* David Schkade et al., *Deliberating About Dollars: The Severity Shift*, 100 COLUM. L. REV. 1139, 1147 n.37 (2000) ("We relied on evidence suggesting that the median judgment of a group of predeliberative individuals is a good predictor of the judgment that group

variation of awards is less for jury awards than for juror awards, thus affecting how and under what circumstances our results regarding juror behavior extend to jury behavior.[103] In spite of this, due to the difficulty of using juries as experimental units, it is very common to use jurors and extrapolate findings to juries. Some authors have simulated jury awards in juror studies by calculating the median award in random groupings of juror awards.[104] We refrained from using this approach for various reasons—in general, because it can frequently lead to misleading results and it relies on a range of questionable assumptions.[105]

Instead, in extending our conclusions regarding juror decision-making to draw conclusions regarding jury decision-making, we rely on our use of various winsorization schemes and various measures of accuracy, spread, and magnitude. Specifically, we analyzed unwinsorized data, data winsorized at the 99th percentile, and data winsorized at the 90th percentile, each using various measures of accuracy, spread, and magnitude, to gain a robust understanding of the effects of interest on various aspects of award distributions—reflecting, for example, various levels of award dispersion. In some ways, this approach has similar effects to the so-called "statistical jury" approach but with fewer negative features.[106]

Furthermore, aside from studies showing that predeliberation juror preferences or certain aggregations of such preferences are good predictors of jury awards, there is a substantial body of literature showing that jury deliberation does not solve the problem of unpredictability.[107]

will reach as a result of deliberation."); Diamond et al., *supra* note 8, at 315-17 (discussing juror and jury damage awards).

[103] Neil Vidmar & Jeffrey J. Rice, *Assessments of Noneconomic Damage Awards in Medical Negligence: A Comparison of Jurors with Legal Professionals*, 78 IOWA L. REV. 883, 897 (1993).

[104] *See, e.g.,* Cass R. Sunstein et al., *Assessing Punitive Damages (with Notes on Cognition and Valuation in Law*), 107 YALE L.J. 2071, 2101 (1998); John Campbell et al., *Countering the Plaintiff's Anchor: Jury Simulations to Evaluate Damages Arguments*, 101 IOWA L. REV. 543, 556-57 (2016); Vidmar & Rice, *supra* note 103, at 897.

[105] *See generally infra* notes 109-112 and accompanying text.

[106] *See generally* Schkade et al., *supra* note 102, at 1171 (comparing "the results of jury deliberation and the results that would be produced by taking the median of nondeliberating . . . groups").

[107] *See* Diamond et al., *supra* note 8, at 314-17; Leebron, *supra* note 8, at 311-16; Schkade et al., *supra* note 102, at 1145-46.

For example, in a study by Diamond, Saks, and Landsman, "the variability did not drop in absolute dollars for jury awards [relative to juror awards] for both economic damages and damages for pain and suffering"; and "[a]s a percentage of mean award . . . jury variability was lower than juror variability for both types of damage awards," but dropped only "from 84% to 78% for economic damages and from 179% to 146% for pain and suffering awards," leaving "substantial unexplained variability . . . across juries."[108] Similarly, Schkade, Sunstein, and Kahneman found that "deliberating juries produce more unpredictability than would be found by taking the median of jurors' predeliberation judgments,"[109] and further, that "27% of non-zero jury dollar verdicts were as high as or higher than that of the highest predeliberation dollar judgment of individuals."[110] Their results led them to conclude that the process of deliberation (relative to grouping juror awards and taking the median) "does not alleviate the problem of erratic and unpredictable individual dollar awards, but in fact exacerbates it."[111] "The safest and most cautious conclusion is that to the extent that unpredictable punitive damage awards raise a serious concern, the problem is not removed by deliberation."[112]

Thus, in light of previous literature and the strong effects observed in our study, as well as the robustness of our findings to various winsorization schemes and measures of accuracy, spread, and magnitude, it is likely that the effects we observed in this study of juror awards would extend to jury awards as well. Nevertheless, "[c]ollective judgments are known to have less variability than individual liability awards," and there is a possibility of "exaggerated . . . effects of anchors";[113] in short, the differences between juror effects and jury effects should be acknowledged and accounted for when considering the implications of our results. We studied juror behavior, and additional inference is required to draw conclusions regarding jury behavior.

[108] Diamond et al., *supra* note 8, at 316-17.

[109] Schkade et al., *supra* note 102, at 1172. The authors also concluded that deliberation causes "dollar awards generally [to] increase, while making high dollar awards substantially increase, in a general severity shift." *Id.*

[110] *Id.* at 1163.

[111] *Id.* at 1139.

[112] *Id.* at 1143.

[113] Campbell et al., *supra* note 104, at 556; *see* Sanders, *supra* note 15, at 494-96 (discussing criticisms based on the disparity between *juror*-based studies and *jury* awards).

Second, our study used *mock* jurors who decided an award following a short summary describing one of two case scenarios rather than real-world jurors who decide an award following an actual trial. In a sense, this study was a simplified "laboratory" experiment aimed at studying juror behavior. A more ideal (although far costlier) experiment would involve an intervention in real-world jury trials. For example, the variability of awards may be exaggerated because mock jurors may treat the situation as hypothetical and ignore certain real consequences of extreme awards, such as bankruptcies, job loss, etc.[114] Note that extreme awards *are* observed in the real world, but likely less frequently (and perhaps less dramatically) than in laboratory experiments. Additionally, our summary descriptions may have affected the variability of awards (both in control and active treatment conditions) relative to real-world trials that involve multifaceted evidentiary support provided to substantiate arguments by the plaintiff and the defendant, and that are presided over by a judge.

As above, however, it is likely that our analyses using various winsorization schemes and various measures of accuracy, spread, and magnitude help to mitigate these effects. Again, our use of winsorization and various estimands allowed us to gain a robust understanding of the effects of interest on various aspects of award distributions. Nevertheless, these award distributions may differ from those that would result from real-world jury trials, which may give rise to unwanted effects.

Third, we used Mechanical Turk to administer the study. Although we took steps to restrict our sample to individuals eligible to serve on a U.S. jury, these steps were not necessarily 100% effective, and, perhaps more importantly, our restricted MTurk population did not necessarily match up perfectly with the adult U.S. population.

There have, however, been numerous studies analyzing the population of MTurk workers. These studies have found that, although there are some differences, the MTurk worker population is relatively representative of the general population—and certainly more representative than traditional pools for surveys and experimentation.[115] Furthermore, we

[114] On the other hand, it is possible that summary descriptions are less emotion-provoking than real-world trials and therefore weigh in the opposite direction.

[115] *See* Adam J. Berinsky et al., *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk*, 20 POL. ANALYSIS 351, 366 (2012) (concluding that "relative to other convenience samples often used in experimental research in political science, MTurk subjects are often more representative of the general population . . . "); Connor Huff & Dustin Tingley, "*Who Are These People?" Evaluating the Demographic*

compared our sample in particular to the population of citizens eligible for jury service and found only modest differences, further reassuring us that our use of MTurk did not distort our results.[116] Nevertheless, as numerous authors have pointed out, the demographics of the MTurk population (and perhaps our sample) may diverge from the general population in a number of respects, including political views, education, and age.[117] Although we believe that these differences are unlikely to have caused any significant distortions, it is important to be aware of them and consider their effects.

Characteristics and Political Preferences of MTurk Survey Respondents, 2015 RES. & POL. 1, 8 (2015) (concluding that "respondents on MTurk are not all that different from respondents on other survey platforms," and that "there are strong reasons for researchers to consider using MTurk to make inferences about a number of broader populations of interest"); Gabriele Paolacci et al., *Running Experiments on Amazon Mechanical Turk*, 5 JUDGMENT & DECISION MAKING 411, 414 (2010) ("Our demographic data suggests that Mechanical Turk workers are at least as representative of the U.S. population as traditional subject pools, with gender, race, age and education of Internet samples all matching the population more closely than college undergraduate samples and internet samples in general."); *see also* Roseanna Sommers, *Will Putting Cameras on Police Reduce Polarization*, 125 YALE L.J. 1304, 1331 n.100 (2016) (addressing concerns regarding MTurk subjects). Berinsky et al. also conclude that MTurk subjects "are apparently . . . not currently an excessively overused pool, and habitual responding appears to be a minor concern," but caution that MTurk users "are notably younger and more ideologically liberal than the public," and they "appear to pay more attention to tasks than do other respondents." Berinsky et al., *supra* note 115, at 366.

[116] Specifically, we compared our Leg I demographic data to population data from the 2010 U.S. Census and identified only two significant differences between demographic averages from our sample and those from the Census. *See generally* American FactFinder, UNITED STATES CENSUS BUREAU, factfinder.census.gov. In particular, regarding sex, 63% of participants in our sample are female, compared to 58% based on the Census data, and regarding age, 2% of participants in our sample are 65 or older, compared to 13% based on the Census data.

[117] *See* John Campbell et al., *Time Is Money: An Empirical Assessment of Non-Economic Damages Arguments*, 95 WASH. U. L. REV. 1, 30 (2017) (addressing concerns regarding MTurk subjects and noting that subjects are "slightly more liberal, educated, young, and wealthy than the population as a whole"). Note also that the jury-eligible population may differ somewhat from the actual jury population.

VI. IMPLICATIONS AND CONCLUSION

The "stark unpredictability" of awards for pain and suffering and punitive damages is arguably unacceptable.[118] We desire "predictability and proportionality,"[119] but, at the same time, require juries to determine these awards through nothing more than a "standardless, unguided exercise of discretion."[120] Awards for pain and suffering and punitive damages should be bound together, in the sense that like cases result in like outcomes. But courts have been unwilling to *bind* such outcomes together actively by using awards in comparable cases as guidance for award determinations.

Recent literature has addressed, theoretically, the major objections to using prior-award information to guide award determinations, and has argued that prior-award information not only reduces the variability of awards, but also improves accuracy, reflecting both bias and variance. The instant study makes a number of important contributions. First, it confirms empirically the hypotheses that prior-award information (whether biased or unbiased) substantially reduces the dispersion of awards and that biased prior-award information causes a distortion in the size of awards in the direction of the bias. These results hold regardless of whether prior-award information is provided in the form of an average, range, or list of prior awards. Additionally, the study suggests a potential for distortion of award size from unbiased prior-award information. Defining award size in terms of a median—the most common measure used in previous literature—we observed no significant distortion caused by unbiased prior-award information. But we observed such distortions using other measures. The effect of unbiased prior-award information on award size is a question for future research.

Most importantly, our study shows that prior-award information causes substantial improvements in the accuracy of awards, and that such improvements are robust to changes in scenario, as well as the bias, variability, and form of prior awards. This means that any introduction of error caused by distorting award size is dominated by the beneficial effects of prior-award information on the dispersion of awards. Furthermore, these effects hold under various definitions of accuracy and various winsorization schemes (simulating, for example, mechanisms by which real-world awards are reduced or capped). Separately, there is

[118] Exonn Shipping Co. v. Baker, 554 U.S. 471, 499 (2008).

[119] Payne v. Jones, 711 F.3d 85, 94 (2d Cir. 2013).

[120] Jutzi-Johnson v. United States, 263 F.3d 753, 759 (7th Cir. 2001).

strong evidence that prior-award information has the beneficial effect of reducing the rate of outlier awards.

These findings provide strong evidence in support of proposals to provide jurors with prior-award information as guidance for determining awards for pain and suffering and punitive damages. Specifically, they provide evidence in support of the behavioral assumptions underlying such proposals, and for the proposition that prior-award information would improve accuracy, even under a wide range of bias, variability, and form conditions. As discussed in Part II, the primary objection to prior-award information is that its benefits rely on the actual comparability of the prior cases to the subject case. But the instant study shows that this is not so: prior-award information improves accuracy even when the prior cases are *not* comparable to the subject case and even when they entail wide-ranging facts and wide-ranging awards, or, for that matter, narrow facts and narrow awards, and regardless of whether the prior awards are provided as an average, range, or list.[121] This effect is not unconditional.

[121] An in-depth analysis of the relative disutilities of unpredictability, on the one hand, and bias or award dependence, on the other hand, is beyond the scope of this article. We note, however, that although there is good reason to be highly resistant to certain types of bias—such as bias based on race or gender—such resistance should arguably not apply to distortions of award size generally, and distortions caused by prior-award information in particular. In the language of this article, we are concerned with the distance between an actual award and a correct award, not whether the source of that distance is bias or variance. Consider, for example, a hypothetical in which a damages award will equal its correct award, say $5 million, on average (i.e., the award is unbiased), but that it is highly variable around its correct award, having a plausible range of $0 to $10 million. A procedure or evidentiary rule that reduces the plausible range of the award to $5 million to $6 million may be much preferred over the status quo, even if such a procedure or rule causes some bias, for example, a new average award of $5.5 million. Furthermore, prior-award information may in fact *reduce* race-based and gender-based bias, and other sources of bias to which courts should be highly resistant, by increasing the relative influence of factors that *are* relevant through a court's selection of prior cases. Similarly, although there is good reason to ensure that certain aspects of a claim's adjudication remain independent of influence from other claims, allowing the outcome of a claim to be influenced by the outcome of another claim is not necessarily contrary to current values, and it may in fact advance them. Claim dependence facilitates the goal that like cases receive like outcomes, and that "more extreme" claims receive "more extreme" awards, etc. It is also noteworthy that, in many contexts, courts do indeed allow the outcomes of claims to be influenced by the outcomes of other claims, notwithstanding the

Some level of bias—for example, bias resulting from extreme non-comparability—would cause a reduction in accuracy. But the effects we observed occurred under relatively extreme levels of bias and variability, and under a range of winsorization schemes and measures of accuracy. Given a reasonable method of selecting prior cases and providing award information to jurors, prior-award information is highly likely to improve the accuracy of awards for pain and suffering and punitive damages.

Furthermore, our results may have implications for alternative methods of controlling award unpredictability, such as damage caps. As highlighted in Part II, instituting damage caps is a highly controversial method of controlling unpredictability that caps damage awards, or certain types of damage awards, without consideration of the particular facts of a case. Notwithstanding frequent criticism on fairness and constitutional grounds, and the transfer of discretion away from jurors, damage caps remain a primary method of controlling unpredictability.[122] Our results provide support for the argument that crude and drastic cap-based methods are not necessary for controlling unpredictability—that prior-award information may constitute a more effective tool for achieving goals of predictability, while maintaining juror discretion and case-by-case adjudication.[123]

Finally, although our finding that prior-award information caused an increased likelihood that participants would provide an explanation for their awards is subject to interpretation, one plausible possibility is that prior-award information provides context for an award, allowing for a more thoughtful or analytical award determination. It is possible that prior-award information provides jurors with a framework in which to make a reasoned, articulable determination, rather than one fitting of Judge Posner's description as standardless and unguided.[124] The meaning of these results is a fruitful topic for future analysis.

In summary, our research provides substantial support for the use of prior-award information, and specifically for the argument that prior-award information is an effective method of reducing the unpredictability

risk of introducing bias and inter-claim dependence. Consider, for example, courts' use of additur and remittitur, comparability analysis in takings cases and other civil contexts, and "proportionality review" and other forms of sentence comparisons in the criminal context.

[122]. *See supra* Section II.A.

[123]. Our results may have implications for other legal contexts as well—for example, criminal sentencing.

[124]. *Jutzi-Johnson*, 263 F.3d at 759.

of awards for pain and suffering and punitive damages while preserving the discretion of the trier of fact and improving the accuracy of awards generally.