Improving the Accuracy of Civil Damage Awards With Claim Aggregation

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Bavli, Hillel J. 2017. Improving the Accuracy of Civil Damage Awards With Claim Aggregation. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:41142054">http://nrs.harvard.edu/urn-3:HUL.InstRepos:41142054</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Improving the Accuracy of Civil Damage Awards with Claim Aggregation

A dissertation presented

by

Hillel J. Bavli

to

The *ad hoc* Committee for the Ph.D. Program in Statistics in Law and Government

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics in Law and Government

Harvard University

Cambridge, Massachusetts

January 2017
Improving the Accuracy of Civil Damage Awards with Claim Aggregation

Abstract

A legal proceeding can be understood as an estimation problem, where the quality of the legal outcome depends on the information used to adjudicate it. This Ph.D. Dissertation is a compilation of five papers that examine a set of methods for improving the information used to adjudicate legal outcomes in the context of civil claims for damages. The papers analyze the possibility of improving civil damage awards by aggregating information across different claims—first, across claims in the class-action context, and second, across claims in altogether separate cases over time. The papers examine such methods statistically, analyzing whether, and under what conditions, aggregating information regarding damage awards or facts in separate claims would improve the quality of awards. And they examine such methods legally, assessing whether, and under what conditions, such methods are permissible under evidentiary, procedural, and constitutional frameworks. The final paper reports and interprets results from a factorial experiment designed to test certain conclusions relating to these methods. In summary, the papers demonstrate that the methods discussed are capable of improving damage awards substantially, and are generally permissible under the law.
TABLE OF CONTENTS

INTRODUCTION 1

I. AGGREGATING FOR ACCURACY: A CLOSER LOOK AT SAMPLING AND ACCURACY IN CLASS ACTION LITIGATION 4

II. SAMPLING AND RELIABILITY IN CLASS ACTION LITIGATION 49

III. THE LOGIC OF COMPARABLE-CASE GUIDANCE IN THE DETERMINATION OF AWARDS FOR PAIN AND SUFFERING AND PUNITIVE DAMAGES 68

IV. SHRINKAGE ESTIMATION IN THE ADJUDICATION OF CIVIL DAMAGE CLAIMS 114

V. GUIDING JURORS WITH PRIOR-AWARD INFORMATION: A RANDOMIZED EXPERIMENT 152
Introduction

A legal proceeding can be understood as an estimation problem. Imagine that there is some “correct” outcome, or at least a range of “acceptable” outcomes, that reflects the policy objectives (such as fairness or deterrence) underlying the substantive law governing the case at hand, as well as the “true” state of the world, including the law and the circumstances of the case. A proceeding, such as a hearing or trial, is necessary, because there is imperfect information regarding the law and the circumstances of the case. The proceeding generates information through, for example, argumentation and the presentation of evidence. And based on this limited information, it produces a legal outcome—an estimate of the “correct” outcome.\(^1\) The quality of the legal outcome, therefore, depends on the information used to adjudicate it.

This Ph.D. Dissertation is a compilation of five papers, corresponding to the five sections below, that examine a set of methods for improving the information used to adjudicate legal outcomes in the context of civil claims for damages. Specifically, the papers herein analyze the possibility of improving civil damage awards by aggregating information across separate claims. Such methods give rise to a range of statistical and legal questions. Statistically, the papers examine whether, and under what conditions, aggregating information regarding damage awards or facts in separate claims would improve the quality of awards. Legally, the papers assess whether, and under what

---

\(^1\) See infra § I.
conditions, such methods are permissible under evidentiary, procedural, and constitutional frameworks.

The first paper, *Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation* (published in *Law, Probability & Risk*),\(^2\) discusses aggregation procedures in the class action context, whereby damage awards for class claims are extrapolated from a small sample of representative adjudications. It examines the conditions under which such procedures would improve the quality of awards, as well as relevant legal considerations. The second paper, *Sampling and Reliability in Class Action Litigation* (published in *Cardozo Law Review de novo*),\(^3\) is a short paper that explains the conclusions in the first paper in non-mathematical terms and extends them in light of the recent Supreme Court decision in *Tyson Foods, Inc. v. Bouaphakeo*.\(^4\)

The third paper, *The Logic of Comparable-Case Guidance in the Determination of Awards for Pain and Suffering and Punitive Damages* (forthcoming in the *University of Cincinnati Law Review*),\(^5\) analyzes a set of aggregation methods applicable in the individual-claim context. Whereas the first set of aggregation methods applies to the class action context, and therefore involves a predefined class of claims arising from common facts and issues, these methods involve the sharing of information across altogether separate cases over time. Specifically, the paper examines the use of “comparable-case


\(^{4}\) 136 S. Ct. 1036 (2016).

guidance”—information regarding awards in prior comparable cases provided to the trier of fact as guidance for determining damage awards—as a method for improving the quality of awards for pain and suffering and punitive damages. The paper argues that such methods are important and permissible under the law, and it applies basic statistical modeling and techniques to explain how and when such methods can be expected to improve damage awards. The fourth paper (coauthored with Yang Chen), *Shrinkage Estimation in the Adjudication of Civil Damage Claims* (forthcoming in the Review of Law & Economics), assumes that the comparable-case guidance methods discussed above could be applied formulaically (or, equivalently, assumes that the trier of fact would incorporate the information according to our model) and derives a range of statistical results regarding the conditions under which such methods would improve the quality of damage awards.

Finally, the fifth paper (coauthored with Reagan Rose), *Guiding Jurors with Prior-Award Information: A Randomized Experiment*, reports and interprets the results of a factorial experiment designed to test the conclusions in the third and fourth papers—and specifically, to test the behavioral assumptions in those papers and whether comparable-case guidance improves the quality of awards (as defined in the papers below) under a robust range of conditions.

In summary, the papers demonstrate that the methods discussed are capable of improving damage awards substantially, and are generally permissible under the law.

---


7 Hillel J. Bavli & Reagan Rose, *Guiding Jurors with Prior-Award Information: A Randomized Experiment*.

8 Note, the papers have been modified from their original form for purposes of consistency.
I. **Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation**

1. **Introduction**

Professor Charles Alan Wright remarked that “[u]nless we can use the class action and devices built on the class action, our judicial system is not going to be able to cope with the challenges of the mass repetitive wrong.”¹ Indeed, US courts increasingly grapple with the limits of aggregation procedures in managing today’s litigation needs. Among the most controversial of such procedures is the use of statistical sampling, and in particular, representative “bellwether” trials, to extrapolate aggregate proof of liability and damages for a class of unlitigated claims.

In 1990, with federal asbestos filings averaging 1,140 per month—double the rate at which such filings were being resolved—U.S. courts searched for novel legal procedures

---

*This article originally appeared in 14 LAW, PROBABILITY & RISK 67 (2015). Citation: Hillel J. Bavli, *Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation*, 14 LAW, PROBABILITY & RISK 67 (2015). It may have been modified from its original form for purposes of consistency. In developing the ideas presented in the current article, I benefitted immensely from the advice and guidance of Professors Donald Rubin and Alan Zaslavsky, and from the thoughtful comments of an anonymous reviewer. Other major contributors, to whom I owe particular gratitude, include Peng Ding, Ruobin Gong, Carl Morris, Joe Blitzstein, Kathryn Spier, David Rosenberg, Louis Kaplow, and the students of Donald Rubin’s graduate seminar course at Harvard. I owe thanks to the Department of Statistics at Harvard University for its support and to the Law, Probability & Risk Editorial Board for their helpful comments and edits.

capable of handling the influx of claims while maintaining procedural due process and ultimately a just resolution of the claims. In this context, controversy stirred over the use of aggregation mechanisms, such as sampling, to resolve the claims.

In Cimino v. Raymark, a 1990 asbestos class action, Judge Parker of the Eastern District of Texas approved a trial plan involving a random sample of 160 claimants from a total of 2,298 class members. He divided the population of class members into five disease categories based on the claimants' injury claims. He then took a random sample from each category to form a representative 160-claimant sample group. The damages claims of each randomly sampled class member were submitted to a jury, and each member of the sample group was awarded his individualized verdict. Then, the average verdict in each disease category was computed and applied to the non-sampled members of each category, respectively. The Fifth Circuit ultimately rejected the sampling procedure, ruling that it violated the defendant's Seventh Amendment right to individualized determinations of whether the products at issue caused plaintiffs' injuries, as well as its right to have jury

---


3 An ad hoc committee, created by the Chief Justice of the U.S. Supreme Court to examine potential solutions to the enormous backlog caused by the asbestos litigation, found that the problem had “reached critical dimensions,” and that the courts are ill-equipped to handle the situation. Id. at 816 (citing Judicial Conference of the United States, Report of the Judicial Conference Ad Hoc Committee on Asbestos Litigation, at 2 (1991)). Among the many recommendations made in its report, the Committee proposed the possibility of sampling asbestos claims for litigation and then extrapolating the outcomes of the remaining claims from those of the sampled claims.


5 See id. at 652-54.
determinations regarding the “distinct and separable issues of actual damages suffered by each extrapolation plaintiff.”

The sentiment in the courts and in the literature—then and now—is that the use of sampling procedures serves purposes of efficiency, but only at the cost of fairness, or accuracy; and that the “benefits of efficiency can never be purchased at the cost of fairness.”

But in a 1992 Stanford Law Review article, Professors Michael Saks and Peter Blanck sought to address the constitutional and reliability concerns that courts and commentators have voiced in opposing the use of sampling. They did not simply argue that properly implemented aggregation can provide the same level of reliability as individualized litigation. Rather, they claimed that in fact “aggregation adds an important layer of process which, when done well, can produce more precise and reliable outcomes”—that “the procedural innovation of aggregation provides a quality of justice that surpasses what courts have, until now, been capable of in any kind of case.”

---


7 Malcolm v. National Gypsum Co., 995 F.2d 346 (2d Cir. 1993). See In re Chevron U.S.A., 109 F.3d 1016, 1023 (5th Cir. 1997) (Jones, J., concurring) (“as Judge Higginbotham cautioned in In re Fibreboard Corp., there is a fine line between deriving results from trials based on statistical sampling and pure legislation. Judges must be sensitive to stay within our proper bounds of adjudicating individual disputes. We are not authorized by the Constitution or statutes to legislate solutions to cases in pursuit of efficiency and expeditiousness.”). In Hilao v. Estate of Marcos, a controversial decision in which the Ninth Circuit approved of the use of sampling to resolve the damages claims of approximately 10,000 claimants, the court nevertheless emphasized that “[t]he statistical method used by the district court obviously presents a somewhat greater risk of error in comparison to an adversarial adjudication of each claim.” 103 F.3d 767, 786 (9th Cir. 1996).

8 Saks & Blanck, supra note 2, at 815-16.
Since *Cimino*, the need for effective aggregation procedures has only intensified. But assertions by Saks and Blanck, as well as other authors, that sampling can increase accuracy as well as efficiency, have been met with skepticism and a general unwillingness to rely on such assertions in ruling on proposed aggregation procedures. Instead, courts have continued to reject sampling on the grounds that accuracy should not be sacrificed for the sake of efficiency. Indeed, although courts and commentators have often rejected sampling explicitly on grounds of protecting the defendant’s constitutional and procedural rights,9 Saks and Blanck are likely correct in remarking that “a major—perhaps the major—due process concern in an aggregated trial is the validity of the outcome.”10 As Saks and Blanck emphasized, and courts have suggested, “The main argument against trial by aggregation and sampling asserts that such trials cannot give the parties as accurate a result as they would obtain through traditional bilateral trials.”11


11Saks & Blanck, supra note 2, at 833. See also *In re Chevron U.S.A, Inc.*, 109 F.3d at 1020 (“our procedural due process concerns focus on the fact that the procedure embodied in the district court’s trial plan is devoid of safeguards designed to ensure that the claims against Chevron of the non-represented plaintiffs as they relate to liability or causation are determined in a proceeding that is reasonably calculated to reflect the results that would be obtained if those claims were actually tried.”).
Arguably, the primary source of such skepticism, and the unwillingness of courts and commentators to rely on assertions that sampling can increase accuracy as well as efficiency, is the generality with which such assertions have been made. In particular, such assertions have not been developed formally, and the literature has not examined the effect of sampling on accuracy, in any rigorous form, under conditions applicable to real-world class actions. Indeed, in most circumstances, sampling does not increase or decrease accuracy categorically; rather, as explained below, its capacity to increase accuracy is a function of the parameters associated with a particular class of claims. Therefore, a deeper understanding of the conditions under which sampling will increase accuracy is necessary.

In the current article, I introduce a framework for examining the conditions under which sampling in the context of class action litigation can increase accuracy in the law. In particular, I develop a model for studying the effects of sampling on accuracy and for deriving the optimal sample size, given a set of parameters associated with a class of claims. I then introduce methods for estimating class parameters and for increasing the utility of claim aggregation.

I begin in Section 2 by motivating the problem with a brief overview of sampling in the context of class action litigation, and a discussion regarding the importance of understanding the conditions under which sampling can increase accuracy. In Section 3, I discuss in further detail the meaning of accuracy and also the aspect of the law that underlies the accuracy-enhancing benefits of sampling: judgment variability. In Section 4, I introduce a basic framework for examining the effects of sampling on accuracy, and for
deriving the optimal sample size, given class parameters and specified constraints. In
Section 5, I discuss a number of important extensions, including possibilities for parameter
estimation and the use of sequential sampling, stratification techniques that may allow for
the reduction of claim heterogeneity and thereby increase the utility of claim aggregation,
and issues related to incentives to initiate claims and to settle under circumstances of claim
aggregation. In Section 6, I conclude.

2. Aggregate Proof and Sampling in Class Action Litigation

A class action is a legal procedure by which multiple (often many) legal claims are
aggregated into a single case for joint determination. Class actions constitute a very
substantial proportion of claims in the USA. Efficiency is the primary purpose of the class
action procedure, without which courts would be overwhelmed with claims brought for
individualized litigation (leading to extreme delays, etc.), and many claimants with
legitimate claims would be unable to afford bringing their claims individually in the first
instance.

Before a court will approve, or “certify,” a proposed class for this form of litigation,
the class representatives must prove to the court that the claims involve common
questions of fact or law, and that it makes sense to try the claims together as a class action
rather than individually. The representatives must show that the claims are sufficiently
numerous to justify the class action procedure; that the claims in the proposed class are
sufficiently common; that the representative claims are typical of the remaining claims; and
that the representatives can adequately litigate these claims on behalf of the class. In
addition, the proposed class must satisfy one of three other requirements—frequently, that common issues predominate over individual issues.

It is an arduous process to achieve class certification. In a sense, it is a trial in itself to determine whether a proposed class should be certified. After all, if the class is not certified, most claimants are unlikely to bring their claims individually; and if the class is certified, the class action has the potential to bankrupt even large companies, and the claimants will have enormous power over the defendant(s) in eliciting a favorable settlement.

An aspect of class certification that has proven particularly difficult for plaintiffs is the requirement that a plaintiff class prove liability and damages on a classwide basis. As the Supreme Court held in *Wal-Mart v. Dukes*, what is important in determining whether to certify a class is not simply the raising of common questions, but the capacity of the class proceeding to produce common answers central to the litigation. Importantly, an individual plaintiff may not escape the requirements of proving liability or injury merely by joining a class. Plaintiffs must be able to show that the defendant is liable with respect to each claimant, and that the defendant’s conduct caused injury with respect to each claimant.

It is here, in the context of proving classwide liability and damages, that class plaintiffs have attempted to employ sampling procedures and bellwether trials.\(^\text{12}\)

\(^{12}\)“Bellwether trials,” or “test cases,” have been used in various legal contexts—such as antitrust and mass-tort litigation—and for a range of purposes. As discussed further below, however, courts have been reluctant to approve such procedures for purposes of determining liability and damages. Arguably, prior to a string of negative court rulings in and after 2011, there was rising momentum for the use of, or at least for further exploration of, such procedures in the mass-tort context. See generally Laura E. Ellsworth & Charles H. Moellenberg, Jr., *Bellwether Trials*, 2 Business and Commercial Litigation in Federal Courts §14:47, 3d ed. (2013); Manual for Complex Litigation §22.315, *Test Cases*, Federal Judicial Center, 4th ed. (2004); Eldon E.
Sampling-based claim aggregation has taken two general forms. The first form involves aggregating and sampling evidence prior to trial—e.g., evidence that the defendant’s behavior caused the harm alleged by the plaintiffs—followed by the presentation of the sampled evidence to a jury, which will then determine an outcome for the entire class. For example, in *Hilao v. Estate of Marcos*, a class action seeking damages for human rights abuses, the district court used a random sample of 137 claims, divided into categories (torture, summary execution and disappearance), from a total of approximately 10,000 claims, to determine appropriate compensatory damages. The district court appointed a special master to supervise the depositions in the Philippines of the 137 sampled claimants and their witnesses. The special master then reviewed the claim forms and the depositions, evaluated the testimony based on instructions provided by the district court, and made recommendations regarding the validity of the claims. He recommended that 6 of the 137 claims in the sample be found not valid; and, for the

---


14 103 F.3d 767, 782-84 (9th Cir. 1996).
remaining 131 valid claims in the sample, he made damages recommendations for each of the three categories. Finally, he made recommendations on damage awards for the remaining (non-sampled) class members based on his recommendations for the sampled claims. In particular, based on his finding that 6 of the 137 claims (4.37%) were invalid, he recommended a 5% “invalidity rate” for the remaining claims, and he recommended an award for the class determined by multiplying the number of claims in each subclass, after discounting by the invalidity rate, by the average award recommended for the randomly sampled claims in each subclass, respectively. A jury trial on compensatory damages was then held. An expert testified regarding the validity of the sampling methodology and the use of inferential statistics, and testimony from the 137 sampled claimants and their witnesses was introduced. Then, the special master testified regarding his recommendations, and his report was provided to the jury. The court instructed the jury that it could accept, modify or reject the recommendations, and that it can reach its own judgment as to damages on the basis of the evidence of the sampled claimants. After deliberating for 5 days, the jury found against only 2 of the 137 claimants in the sample, but generally (with some exceptions) adopted the special master’s recommendations. The district court then entered judgment for 135 of the 137 claimants in the amounts the jury awarded, and entered judgment for the remaining claimants in each of the three subclasses in the amounts the jury awarded, to be divided pro rata.\textsuperscript{15}

\textsuperscript{15} Id.; see also Blue Cross and Blue Shield of New Jersey, Inc. v. Philip Morris, Inc., 178 F. Supp. 2d 198 (E.D.N.Y. 2001) (Individual action by health insurer against tobacco companies in which a sample of 156 deponents was selected, and statistical evidence was presented to the jury in combination with studies, relevant portions of videotaped deposition testimony, and expert testimony to enable valid statistical inference to prove causation-to determine whether smokers were in fact deceived by statements made by the tobacco industry); Walker &
The second form—the focus of the current article—involves the use of representative “bellwether trials” to resolve classwide issues. The procedure approved by Judge Parker in Cimino v. Raymark is illustrative. In Cimino, a class of asbestos claims was divided into five disease categories, and a small random sample of claims was selected from each category for individualized jury determination.\(^\text{16}\) The court awarded the sample-group claimants their respective individualized determinations, and then the average award in each disease category was computed and awarded to the remaining (non-sampled) claimants in each category, respectively.\(^\text{17}\)

Consider the important case Wal-Mart v. Dukes,\(^\text{18}\) a class action lawsuit by approximately 1.5 million current and former female employees against Wal-Mart alleging that the super chain store discriminated against them on the basis of gender. The plaintiffs argued that, in order to show classwide liability, they must show that Wal-Mart indeed discriminated against women, and that the 1.5 million class members were females who were employed by Wal-Mart during the relevant period. They argued that this issue-

---

\(^{16}\) Methods for selecting bellwether trials can be categorized as follows: “(i) plaintiffs’ counsel selects the sample; (ii) each side selects an equal number of sample cases; (iii) the court selects the sample from nominees proposed by the parties; or (iv) the samples are randomly selected from agreed-on categories of claimants.” 2 McLaughlin on Class Actions 8:6 (10th ed.). See Sterling v. Velsicol Chemical Corp., 855 F.2d 1188, 1196 n.6 (6th Cir. 1988). Hereafter, unless stated otherwise, the reader should assume that claims are selected for the sample group randomly from the class, or from class strata if the class is stratified, such as in Cimino.

\(^{17}\) See generally Cimino, 751 F. Supp. at 652-54. Note, although Cimino involved sampling as a means of resolving damages issues, herein I consider sampling in a broader context—including its use to resolve issues of liability.

\(^{18}\) 131 S.Ct. 2541, 180 L.Ed.2d 374 (2011).
whether Wal-Mart discriminated against women-is the common question that binds all members of the class, and that makes a class action, rather than individualized litigation, appropriate.

The defense countered that, as a matter of law, if an employer is sued on grounds of discrimination then irrespective of whether the employer was discriminatory or not, the employer may put forth the defense that the particular conduct alleged by a given plaintiff was caused by legitimate, rather than discriminatory, purposes (e.g., that a particular employee stole from the cash register and was fired on those grounds alone). It argued that this affirmative defense must apply on an individualized basis, rather than through some shortcut that may be applied commonly to the class as a whole.

The lower courts approved a trial plan that was intended to overcome the problem described by the defendants. The plan involved taking a small sample of claims (the “sample group”) that would be examined to determine what proportion of the sample would result in a determination of Wal-Mart’s liability. Further, damages would be determined for each of the sampled claims. The claims in the sample group would receive individualized determinations. But the court would then assign damages to the remainder of the class (the “extrapolation group”) by averaging the damage determinations from the sample group (including zero damages for claims in which Wal-Mart was found not liable) and multiplying that average by the number of remaining claims.

The Supreme Court reversed the decision of the lower courts and outright rejected the use of sampling to prove classwide liability and damages where such sampling abridges
the right of a defendant to defend itself on an individualized basis under the substantive law. The Supreme Court viewed the trial plan as a form of “Trial by Formula” and “disapprove[d of] that novel project.”

_Wal-Mart_ reaffirmed and perpetuated a significant body of case law rejecting the use of sampling as a means of offering aggregate proof of classwide liability and damages.\(^\text{19}\)

In the wake of _Wal-Mart_, an understanding of conditions under which sampling can increase accuracy as well as efficiency is of fundamental importance. Professor Wright was likely correct when he warned of the necessity of building on the class action procedure to address the challenges of the mass repetitive wrong.\(^\text{20}\) Arguably, in light of the nature of today’s social and economic interaction, the need for such procedures has never been greater. The courts are nevertheless increasingly unwilling to benefit from the efficiencies of aggregation, but under the false premise that such benefits come only at the cost of accuracy. But perhaps the greatest benefits of aggregation in fact lie in its ability to increase accuracy in addition to efficiency.

As Saks and Blanck suggested, the validity of the case’s outcome is a major (if not _the_ major) source of courts’ antagonism towards sampling, including antagonism expressed in

\(^{19}\) _See, e.g.,_ McLaughlin v. American Tobacco Co., 522 F.3d 215, 224 (2d Cir. 2008); _Cimino_, 151 F.3d at 297; _In re Fibreboard Corp._, 893 F.2d at 706; Basco v. Wal-Mart Stores, Inc., 216 F. Supp. 2d 592 (E.D. La. 2002). _See also_ Espenscheid v. DirectSat USA, LLC, 705 F.3d 770 (7th Cir. 2013) (holding that plaintiffs’ proposed trial plan was not feasible due to variance in damages across class members, thus rejecting plaintiffs’ proposal to address such variance by presenting testimony at trial from 42 representative members of the class). _But see_ Saks & Blanck, _supra_ note 2, at 837 (emphasizing that “from the viewpoint of defendants, even if there are relatively large errors, with numerous over- and under-awards, all of those differences will cancel each other out and the average award will be the same in the collective trial as it would have been with the individualized determinations”). For case law and discussion, see 2 _McLaughlin_ on Class Actions, _supra_ note 16, §8:7.

\(^{20}\) _See supra_ note 1.
terms of courts’ concern for the effect of sampling on a defendant’s procedural rights.\textsuperscript{21} If this is the case, then an understanding of the conditions under which sampling can increase accuracy may lead courts to reverse their current course.\textsuperscript{22}

Further, regardless whether accuracy is at the center of judicial decisions rejecting sampling, an understanding of the accuracy-enhancing benefits of sampling enables courts and lawmakers to better account for the costs of substantive laws that prevent, or that do not enable, such procedures. In fact, currently, numerous federal statutes explicitly or impliedly permit sampling as a tool for proving damages. For example, the Hart-Scott-Rodino Antitrust Improvements Act states that in \textit{parens patriae} actions by state attorneys general in which there has been a determination that a defendant agreed to fix prices, “damages may be proved and assessed in the aggregate by statistical sampling methods, by the computation of illegal overcharges, or by such other reasonable system of estimating

\textsuperscript{21} See Saks & Blanck, \textit{supra} note 2, at 833; \textit{see also supra} notes 10-11, and accompanying text; 2 McLaughlin on \textit{Class Actions, supra} note 16, §8:7. Note, arguably \textit{Wal-Mart} may be viewed as a special case—and arguably distinguishable from other cases in which a plaintiff class seeks to employ sampling to prove liability or damages—in that it was in Wal-Mart’s interest to defend against the plaintiffs’ claims on a claim-by-claim basis, \textit{independent of the validity of the damages incurred by Wal-Mart in the aggregate}. In particular, it was in Wal-Mart’s interest to sort valid claims from invalid claims and thereby set a strong precedent for employees that they will be rewarded based on their performance, and not based on windfall legal damages. Arguably, a court may fulfill this sorting function as well by ordering post-award allocation hearings.

\textsuperscript{22} Note, Saks and Blanck do address the importance of maintaining procedural justice for its own sake, but they consider the procedural \textit{injustice} that can result from traditional procedures for the majority of the class, and they ultimately conclude that, at least with respect to the plaintiffs in actions like the asbestos litigation, aggregation can provide more procedural justice than the alternatives. \textit{See Saks & Blanck, supra} note 2, at 838-39. Saks and Blanck argue that “a closer look at the aggregated trial, at least in the mass tort context, suggests that this procedure does not necessarily violate traditional notions of due process under the Fifth and Fourteenth Amendments.” They assert that, “[i]n fact, the absence of such procedures is tantamount to denying many litigants their due process trial rights altogether.” \textit{Id.} at 826. Further, regardless whether procedural justice is valued in its own right, few would argue that such value can be viewed entirely independently of its instrumental value in ensuring accurate outcomes, or that “procedural justice” should be preserved regardless of the potential costs with respect to accuracy.
aggregate damages as the court in its discretion may permit without the necessity of separately proving the individual claim of, or amount of damage to, persons on whose behalf the suit was brought."

Thus, in the following sections, I develop a framework for examining the conditions under which sampling can increase the accuracy of legal determinations. I begin by describing the concepts of accuracy and variability in the law.

3. **Accuracy and Variability in the Law**

Central to the debate over sampling and to the framework developed herein are the concepts of *accuracy* and *judgment variability*. I begin, in the current section, with a general discussion of the concepts, and then introduce a more formal conceptualization in Section 4 below.

3.1 **The Meaning of Accuracy**

One relatively simple approach to describing accuracy in the law is to assume that, for every legal claim, there exists a “correct” outcome that can, in theory, be determined as a function of the complete facts surrounding a given claim and the “true” state of the law at the time of the claim. In this sense, assuming that complete knowledge of the facts and the

---


24 A more complex approach is to consider a probability distribution of “correct” outcomes associated with each and every legal claim.
state of the law is unavailable, a trial will result in an estimate of the correct outcome.25 As such, each legal outcome involves an error term that, if possible to compute, would describe the distance between the actual outcome and the correct outcome. We can then define accuracy as the proximity of the actual verdict to the correct verdict. 26

This description is consistent with Saks and Blanck’s assertion that it makes sense “to think of the ‘true’ award as the average of the population of possible awards” that would result from trying the same case repeatedly for an infinite number of times under various conditions (e.g., before different juries, by different attorneys, using different permutations of facts, etc.).27 For simplicity, I ignore the potential for bias and adopt this formulation of the “correct” outcome associated with a given claim.28

Suppose, for example, that a cigarette smoker files a lawsuit against a major cigarette manufacturer for failure to adequately inform the consumer of the harms associated with smoking. Following trial, a jury returns a verdict finding the cigarette manufacturer liable and awards the plaintiff damages in the amount of $800,000. We can

---


27 Saks & Blanck, supra note 2, at 833-34.

28 But see infra note 41. Note, if judgment variability were entirely due to variability in the composition of the jury, we could arrive at an intuitive variant of Saks and Blanck’s “true” award by imagining a hypothetical situation in which every possible jury in a given geographic region observes the trial and independently (if hypothetically possible) determines a verdict. Then the “correct” outcome may be described as the average of all such verdicts.
assume that, given complete information regarding the law and the facts of the case, we can compute the correct outcome; but since perfect information regarding the facts and the law is unavailable, the jury arrives at a verdict, which serves as an estimate of the correct outcome. In our example, the accuracy of the verdict is defined as the proximity of the verdict, $800,000, to the correct outcome.29

Thus, as discussed above, accuracy is often understood as a fundamental objective of the law. Saks and Blanck discuss this interest within the context of what they refer to as “distributive justice”; and they highlight the instrumentalist perspective that “the constitutional purpose of the due process clause is to ensure the most accurate decision possible.”30 Further, however, maximizing the accuracy of legal determinations is likely to serve the aims of the law, whether or not accuracy, or distributive justice, is itself the primary objective of the law.31 Whether accuracy is understood as a means of achieving

---

29 It may be more intuitive to consider the concept of a “correct” verdict in the context of a criminal trial. Consider, for example, the O.J. Simpson murder trial. Polls show that over 50% of Americans believe that the jury arrived at an incorrect verdict. Implicit in the public’s disagreement with the verdict is an assumption that there exists a “correct” outcome. The framework described above is intuitive: had the jury known the true facts of the case, and had there been no ambiguity regarding the application of the law to the facts of the case, the jury would have arrived at the correct conclusion. But given ambiguity regarding either the facts of a case or the state of the law, it is unclear whether a criminal defendant in fact satisfied the elements of the crime charged; and the jury must arrive at a verdict—“guilty” or “not guilty”—which serves as an estimate of the correct outcome, “guilty” or “not guilty.” Further, given a finding of guilt, the court must determine an appropriate punishment (e.g., an appropriate prison sentence), which can similarly be understood as an estimate of some correct outcome. Thus, ultimately the court may arrive at a prison sentence that serves as an estimate of some correct sentence, where a finding of “not guilty” translates, for example, to a prison term of zero.

30 Saks & Blanck, supra note 2, at 833 (citing Redish & Marshall, supra note 10, at 476-77).

31 It has been argued, for example, that the law should aim to maximize welfare rather than fairness. See Louis Kaplow & Steven Shavell, Fairness Versus Welfare, 114 HARV. L. REV. 961 (2001). Regardless whether maximizing accuracy is deemed to be a fundamental goal of the law in and of itself, there is a strong argument that accuracy (or at least the perception thereof) is fundamental to achieving alternative goals, such as welfare. For example, a low degree of accuracy in the law may serve to dilute the deterrence effect intended by the law;
important legal objectives, or itself a critical objective of legal procedure, the role of accuracy in the law is fundamental. Further, as emphasized above, concern for accuracy is at least an important factor—if not the primary factor—underlying the widespread rejection of sampling-based aggregation in the context of class action litigation.

Therefore, throughout the article, I assume that, with respect to sampling, accuracy is a fundamental objective of legal procedure. Further, I assume that sampling, in general, promotes goals of efficiency, since sampling-based trial plans are commonly offered as an alternative to individualized litigation. However, I return to the matter of efficiency in Section 5.

3.2 Judgment variability: A Source of Error in the Law

The class action procedure allows a court simultaneously to decide, with binding effect, numerous claims involving common issues. At the cost of foregoing individualized litigation, the class action device offers a range of benefits—primarily, litigation efficiency and economy, and the protection of rights of individuals who may not be able to bring claims on an individual basis.\(^32\)

\(^32\) Am. Pipe & Const. Co. v. Utah, 414 U.S. 538, 553 (1974); Crown, Cork & Seal Co., Inc. v. Parker, 462 U.S. 345, 349 (1983) (stating that the principal purposes of the class action procedure are the “promotion of efficiency and economy of litigation”); Andrews v. Chevy Chase Bank, 545 F.3d 570, 577 (7th Cir. 2008) (emphasizing that the primary purposes of the class action are judicial economy and efficiency); Haley v. Medtronic, Inc., 169 F.R.D. 643, 647 (C.D. Cal. 1996) (“Class actions have two primary purposes: (1) to accomplish judicial economy by avoiding multiple suits; and (2) to protect the rights of persons who might not be able to present claims on an individual basis.”); In re Caesars Palace Sec. Litig., 360 F. Supp. 366, 398 (S.D.N.Y. 1973) (suggesting that the primary purpose of the class action device is “to give small investors a reasonable opportunity to vindicate their claims in a manner which will not place an undue financial burden upon them”).
Sampling has been used as a means of preserving the class action even where plaintiffs are unable to otherwise offer classwide proof of liability and damages. The idea is to determine outcomes through individualized litigation for a small sample of claims, and then to estimate outcomes for the extrapolation-group claims in the form of an aggregate determination extrapolated from the sample-group determinations. Opponents of aggregate proof argue that where classwide liability and damages cannot be shown absent extrapolation from a sample group, the court should reject the class action, e.g., in favor of individualized litigation, wherein each plaintiff must bring her claim individually if at all. That is, opponents of sampling argue that the court should not allow “procedural economy to overshadow substantive law,” it should not sacrifice accuracy for the sake of efficiency.

But opponents of sampling have neglected a critical feature of litigation: as highlighted above, all verdicts are, in a sense, estimates of the truth; and trying one case before two independent juries is highly likely to result in two different estimates of the correct outcome.

In terms of variability, opponents of sampling have focused on the error that results from extrapolation—e.g., from applying the average of the sample-group determinations to

---


the remaining claims. Assuming the sample group is representative of the class as a whole, extrapolation error is a function of claim variability. If all claims were homogeneous, deserving equal awards of damages, then applying the average determination in the sample group over all claims in the extrapolation group would cause minimal error. On the other hand, if the class entails a high degree of claim variability, then applying a single point estimate—such as the average determination in the sample group—to all claims in the extrapolation group will result in a high degree of error. This is a legitimate concern; and individualized litigation is an effective means of addressing such error. That is, regardless the degree of claim variability, if claims are decided individually, there is no extrapolation, and therefore no extrapolation error.

However, opponents of sampling have neglected a different form of variability: even if the claims in a class are (i) entirely homogeneous, and (ii) litigated individually, there would nevertheless be significant variability in the verdicts determined for those claims. I refer to this form of variability as judgment variability. Judgment variability results from a range of sources, including jury composition, variability in the presentation of evidence, variability in the judge assigned to preside over the case, etc.

Judgment variability can be understood as a very significant source of error in the law. For example, imagine a class of 1 million homogeneous claims brought in the context of a class action lawsuit (against, for example, the largest cigarette manufacturers). A jury

---

verdict in the representative action that is merely $1,000 greater or smaller than the correct verdict—a differential caused by judgment variability—would result in a whopping total error value of $1,000,000 * $1,000 = $1 billion.

Courts are in fact willing to accept a degree of error in order to avoid the burdens associated with litigating 1 million claims individually. But the error that courts have acknowledged and accepted in return for such procedural efficiency is that associated with variability in claims, rather than that associated with judgment variability. Indeed, even litigating each of the million claims individually would result in a very large error value, as each of the million claims involves an error term resulting from judgment variability—and so the total error still involves the aggregation of 1 million error terms.

But imagine now a (costly) hypothetical procedure in which each and every claim were litigated 10 times independently, and in which the outcome associated with each claim were computed by taking the average of the 10 verdicts associated with that claim. That is, start by taking the first claim and litigate it before ten independent juries to obtain 10 independent verdicts. Then, assign the average of the 10 verdicts as the

---

36 Of course, if the claims are too heterogeneous, the court will refuse to certify the class. The class certification procedure described above is intended to ensure the representativeness of the litigated claims—in essence some significant degree of homogeneity. To be clear, our current example involves homogeneous claims and therefore zero claim variability. The courts are therefore likely to certify such a class.

37 Let us for now ignore complexities associated with statistical independence.

38 We can imagine a single trial before 10 independent juries; but it is better to imagine the case being litigated anew before each of the 10 juries, since judgment variability arises from variability in litigation (e.g., in the presentation of evidence) as well as jury variability.
outcome of the first case. By applying this aggregated outcome, rather than any single verdict, we may reduce the error resulting from judgment variability to nearly nothing.\textsuperscript{39}

Applying this replication procedure to all million claims would be incredibly costly—perhaps far more costly than can be justified by the benefits of the procedure—but the purpose of the hypothetical is simply to make the point that replication allows us to reduce error associated with judgment variability.

Now, in the context of a single claim, the benefits of the costly replication procedure are relatively low, since the error resulting from judgment variability is low in the first place, as it is the error associated with only a single claim. Similarly, applying replication to each and every claim in the hypothetical set of 1 million claims described above is incredibly costly and impractical; and in fact can be viewed just as we view replication in a single claim, but 1 million times.

Amazingly, however, in the context of a homogeneous class, we do not need to apply replication to each and every claim in order to realize the benefits of such replication. In fact, assuming that the claims are homogeneous, we can apply the replication procedure a single time—either litigating a single claim, \textit{e.g.}, before 10 independent juries, or by litigating 10 claims, each before an independent jury\textsuperscript{40}—and realize the error-reducing benefits of the procedure as though we applied replication for each claim individually!

\textsuperscript{39}See \textit{supra} note 28, and accompanying text.

\textsuperscript{40}These procedures are equivalent, since the claims are homogeneous.
For this reason, the class action context may provide a unique opportunity to reduce error resulting from judgment variability wholesale—that is, without incurring the “full” costs of such procedures. Thus, if replication can be used to mitigate the error resulting from judgment variability, it is possible that aggregation—and sampling procedures in particular—can result in far more accuracy than even the purported ideal of individualized litigation.41

Opponents of sampling have largely neglected or misunderstood the possibility that sampling can reduce judgment variability and thereby increase accuracy in the law. For example, Professor Robert Bone, commenting on Saks and Blanck’s reasoning regarding the utility of sampling as a function of claim variability,42 states that the “problem with [their] argument is easy to see,” remarking that if claims were homogeneous, then sampling would be unnecessary in the first instance.43 The author explains that “[i]f there were perfect homogeneity, there would be no problem with mass tort litigation from the point of view of outcome accuracy. If all cases were identical, then the trial of any one would be just as good as the trial of any other, and the aggregation could be easily adjudicated with a

41 Note, depending on the severity of the judgment variability associated with a given set of claims, aggregation may well enhance accuracy even if replication yields a biased outcome. That is, if we relax the assumption in § 3.1 above, that averaging verdicts from infinitely many repeated trials would yield the “correct” outcome, or if we otherwise assume that aggregation with respect to a particular set of claims is likely to result in a biased outcome—an outcome that does not approach the “correct” outcome as we average more and more independent verdicts—then aggregation may nevertheless enhance accuracy if the reduction in error resulting from judgment variability exceeds the error resulting from the bias.

42 See Saks & Blanck, supra note 2, at 833-37.

simple class action.” But Professor Bone’s reasoning neglects to account for the error associated with judgment variability. Because judgment variability exists irrespective of the degree of claim variability, the error resulting from a simple class action may be far greater than that resulting from a sampling based aggregation approach.

As discussed further below, homogeneity provides ideal circumstances for error-reducing sampling and extrapolation procedures. Further, it is important to understand that, with respect to outcome accuracy, homogeneity does not allow for easy adjudication “with a simple class action,” as Professor Bone suggests. To the contrary, a single class adjudication for a large, even homogeneous, class will result in a judgment-error value that is multiplied over the entire class of claims, as illustrated in the hypothetical above involving 1 million claims against the large cigarette manufacturers.

Of course, even claims in the class action context are not perfectly homogeneous. But the error-reducing benefits of sampling do not necessarily require homogeneity. Indeed, as shown below, whether claim aggregation increases accuracy depends on whether the error resulting from judgment variability dominates the error associated with claim variability or vice versa.45

4. A Framework for Examining Sampling and Accuracy in Class Action Litigation

---

44 Id. at 578 n. 48.

45 See Saks & Blanck, supra note 2, at 833-37.
As suggested in the previous section, replication can serve as a fundamental tool for addressing judgment variability; and the class action context may often allow courts to realize the benefits of replication procedures without incurring the enormous costs such procedures normally entail. In *Cimino*, for example, the court sampled numerous claims within each disease category to arrive at an aggregated determination, which would be applied to the unlitigated claims within each category, respectively. Efficiency was the primary purpose of the procedure; but given a certain degree of homogeneity within each disease category, the aggregated determinations had the potential to be more accurate than the individualized determinations.

Importantly, the court in *Cimino*, and other courts employing aggregation procedures, are constrained with respect to the procedures they may apply. Had Judge Parker, in *Cimino*, believed with confidence that the aggregated determinations were more accurate than the individualized determinations, and had he designed the trial plan with the sole goal of maximizing accuracy—unconstrained by other considerations—he may have decided not only to apply the aggregated determinations to the unlitigated claims, but to apply them to the *litigated* claims as well. That is, based on the assumption that the aggregated determinations are more accurate than individualized determinations, in addition to applying the aggregated determination in a given disease category to the unlitigated claims in that category, Judge Parker may have decided to replace the individualized determinations in the sample group of that category with the category’s aggregated determination—the average damages award in that category. Otherwise, only the
extrapolation group would receive the accuracy-enhancing benefits of replication, while the sample group would suffer a high degree of judgment variability.

Thus, if a court presiding over an action involving a homogeneous class were unconstrained in its determination to maximize accuracy, it may, for example, sample all claims in the class for individualized litigation and then replace all individualized determinations by an aggregated determination.

In fact, a court may continue to reduce judgment variability by replicating beyond the number of claims in the class—for example, by litigating each claim in the class twice. Replication is costly, however, and presumably the court’s desire for accuracy would be constrained by litigation costs.

But moreover, courts presiding over class actions are constrained by legal considerations extraneous to concerns for accuracy or efficiency.

In the current article, I assume that a court’s ability to increase accuracy is constrained in a particular manner: that is, I assume that a court will not replace an individualized determination with an aggregated determination. A court must choose whether to sample a given claim for individualized litigation or to preserve its eligibility to receive an aggregated determination extrapolated from the sampled claims. Thus, if a court samples a claim for individualized litigation and the determination of an individualized verdict, it will not later replace the individualized determination with an aggregated determination, such as the sample-group average.
It is appropriate to impose such a constraint, even assuming a court’s willingness to permit sampling-based aggregation. While there is clear precedent for the implementation of procedures in which a representative claim is litigated and its determination extended to other claims, there is generally no precedent (within the context of sampling or other) for a procedure involving the replacement of an individualized determination with an aggregated determination.

Indeed, there is a fundamental distinction between extending representative determinations to unlitigated claims and replacing individualized verdicts with aggregated verdicts. For example, although some courts have held that a defendant’s Seventh Amendment right to a jury trial is not violated by the former, a procedure in which a court replaces individualized jury verdicts with aggregated verdicts is highly likely to be violative of the Seventh Amendment. As Justice Brandeis explained:

[The Seventh Amendment] does not prohibit the introduction of new methods for determining what facts are actually in issue, nor does it prohibit the introduction of new rules of evidence. Changes in these may be made. New devices may be used to adapt the ancient institution to present needs and to make of it an efficient instrument in the administration of justice. ... Indeed, such changes are essential to the preservation of the right. The limitation

---

46 In addition to sampling procedures, the class action device itself arguably provides for a form of such treatment.

47 See, e.g., In re Estate of Marcos Human Rights Litigation, 910 F. Supp. 1460, 1468-69 (D. Hawai‘i 1995) (“Here, the jury did determine the facts of the case, as the substance of the action was presented to the jury. There would be no benefit to either side in having the entire class testify given the repetition in the claims. Rule 23 of the Federal Rule[s] of Civil Procedure does not mandate the presence of each member of the class. Therefore, by choosing a random sample of 137 claimants in an aggregate trial, neither side was deprived even the form of their right to a jury trial.”).

imposed by the amendment is merely that enjoyment of the right of trial by jury be not obstructed, and that the ultimate determination of issues of fact by the jury be not interfered with.\textsuperscript{49}

Whereas there is precedent for extending representative verdicts to a class of claims, \textit{replacing} individualized verdicts with aggregated determinations arguably substitutes the jury’s power to determine issues of fact with a determination arrived at by the court’s aggregation procedure.

Similar reasoning may apply with respect to due process concerns as well. Even Judge Parker—who, notwithstanding ultimate reversal by the Fifth Circuit, strongly sided with the legality of sampling procedures in \textit{Cimino}—implemented a plan that highlighted the importance of allocating individualized awards to the sampled claimants, in addition to ensuring “representativeness” of the sampled claims. Arguably, these features of the trial plan were critical to distinguishing the sampling procedure in \textit{Cimino} from the “lump sum” approach rejected by the Fifth Circuit in \textit{In re Fibreboard Corp.}\textsuperscript{50}

Further, problems with verdict replacement are exacerbated by the fact that a court will not know with certainty whether to attribute observed outcome variability to judgment variability or to claim variability, and there is no precedent for attributing observed outcome variability to the former rather than the latter. Indeed, due process, the Seventh Amendment right to jury determination, and current notions of procedural justice

\textsuperscript{49} Blue Cross and Blue Shield of New Jersey, Inc. v. Philip Morris USA Inc., 344 F.3d 211, 225-26 (2d Cir. 2003) (quoting \textit{In re Peterson}, 253 U.S. 300, 309-10 (1920)).

\textsuperscript{50} 893 F.2d 706 (5th Cir. 1990). See \textit{Cimino}, 751 F. Supp. at 664-66; Saks & Blanck, \textit{supra} note 2, at 823-24. Thus, if the procedure in \textit{Cimino} had not ultimately been rejected by the Fifth Circuit, the court would have required individualized allocation of damage awards to sampled claimants nevertheless.
likely require courts to attribute variability in verdicts to the latter, and similarly, likely preclude the possibility of replacing an individualized determination with an aggregated one—even assuming courts’ willingness to sample and apply an aggregated determination to unsampled claims. For example, if a jury awards 100 distinct damage awards to 100 sampled claimants, the Seventh Amendment likely precludes the assumption that such variability is due to judgment variability, rather than to the jury’s finding of differences among the claims, and it likely forbids the replacement of such awards with an aggregated determination.\footnote{See generally \textit{Blue Cross and Blue Shield of New Jersey}, 178 F. Supp. 2d at 256 (“The historical record demonstrates that the Framers’ main objective in drafting the Seventh Amendment was to limit the ability of an appellate court to overturn a civil jury’s finding of fact. There is no indication they intended to constrain the trial judge’s substantial discretion to employ appropriate procedural mechanisms in managing a trial so as to arrive at the truth—or as near to the truth as time and humankind’s limitations allow.” (citing \textit{Simon v. PhilipMorris, Inc.}, 200 F.R.D. 21, 33 (E.D.N.Y. 2001))). Note that courts that have approved sampling-based trial plans have consistently permitted extrapolation only with respect to the unlitigated claims—those in the extrapolation group—while assigning the respective individualized determinations to claims in the sample group. \textit{See, e.g., Hilao}, 103 F.3d at 782-87; \textit{Cimino}, 751 F. Supp. at 652-54. It is often unclear whether courts have done so due to the possibility that they have treated sampling as a second-best option, to be applied only when individualized litigation is not practicable, or because they have presumed that replacing an individualized verdict with an aggregated determination is precluded by the Constitution and rules of procedure. But numerous courts and scholars have emphasized the advantages of aggregation over individualized litigation in the complex-litigation context. \textit{See Cimino}, 751 F. Supp. at 659-66. \textit{See generally \textit{Blue Cross and Blue Shield of New Jersey, Inc.}}, 178 F. Supp. 2d at 248, 258 (finding that “[w]hen, as in the case at bar, the plaintiff is an entity that has suffered its injury in the aggregate, statistical evidence is notably a more accurate and comprehensive form of evidence than would be the testimony of millions of smokers”; and emphasizing that “[m]any authorities demonstrate that jury fact finding is enhanced by the use of aggregating techniques” and that “[t]he essential functions of the Seventh Amendment are enhanced, not limited, by procedures which streamline and focus jury fact-finding.”) Further, for the reasons above, it is likely that regardless whether courts view sampling as a second-best strategy, they would be unwilling to replace an individualized outcome with an aggregated one.}

51
ambitions to reach higher and higher degrees of accuracy. The constraint implies the existence of an optimal sample size, with respect to accuracy, as opposed to conditions under which a court can continue to increase accuracy by increasing replication.

Thus, to introduce a framework for analyzing the circumstances under which aggregation can increase the accuracy of legal outcomes, I will assume conditions that can be described as “reductive sampling.” To illustrate, suppose we have a pond of sick fish, and our goal is to save as many fish as possible. We need to “sample” fish, and thereby perform tests on them, in order to determine the nature of the disease and how to treat it. Now assume (i) that as we increase the number of fish we sample, we increase the accuracy of our information regarding the nature of the disease and how to treat it; but (ii) that “sampling” a fish requires that we kill it in order to perform tests on it, etc. Thus, we are presented with a tradeoff: the more fish we sample, the more accurate our information regarding the disease and how to treat it, but since sampling a fish involves killing it, the more fish we sample, the fewer fish remain to treat. If we sample no fish, they will all eventually contract and die of the disease. On the other hand, if we sample all of the fish in the pond, we will have a high degree of knowledge of the disease and how to treat it, but we will have no fish remaining to treat! Thus, assuming circumstances described by reductive sampling, in which sampling a unit reduces, or (as here) altogether destroys, the value obtained from later extrapolation with respect to that unit—for example, treating or otherwise applying an operation to the unit—leads to the emergence of an optimization
problem: how many fish must we sample so as to balance our interests in learning how to treat the disease on the one hand and preserving fish to treat on the other.

By sampling, we partition the population of claims into (i) a sample group, in which claims are litigated and verdicts determined on an individualized basis; and (ii) an extrapolation group, in which claims receive the average of the determinations in the sample group. The model therefore highlights the tradeoff between the accuracy of the sample mean, on the one hand, and the number of claims that remain in the extrapolation group, on the other.

In earlier sections, I have highlighted two forms of variability: claim (i.e., fact) variability and judgment variability. Claim variability might include, for example, various disease types suffered by smokers in a class suing a cigarette manufacturer. Judgment variability, on the other hand, refers to the variability in the verdicts, holding the facts across claims constant. As mentioned above, judgment variability might result from variability in the presentation of evidence, variability in the composition of a jury and variability in the selection of the judge to preside over the case.

Let us begin by assuming that judgment variability is the sole source of variability. That is, assume that claims are perfectly homogeneous. This is a good starting point for analysis for two reasons. First, we may in fact observe a significant degree of claim homogeneity in the context of class action lawsuits. Moreover, a class may be subclassified for the purpose of creating a high degree of homogeneity within each subclass. Second, assuming zero fact variability allows examination of the full theoretical benefit of
aggregation before addressing the problem of balancing concerns of judgment variability with those of fact variability.

4.1 Homogeneous Claims

Assume we have a class of $N$ homogeneous claims from which we sample $n$ claims. Let $\mu$ be the true mean of the verdicts over the total population of possible verdicts—i.e., the “correct” verdict. Let $X_i$ be a random variable defined by the actual verdict in the $i$-th claim, for $i = 1, 2, 3, ..., n$, where $X_i \sim N(\mu, \sigma^2)$ i.i.d. Further, define $\bar{X}_n$ as the mean of the $X_i$, for $i = 1, 2, 3, ..., n$, where $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Then, pursuant to the above description, let us represent the total error associated with all $N$ claims as the following sum of squares residual:

$$R = \sum_{i=1}^{n} (X_i - \mu)^2 + n(N - n)(\bar{X}_n - \mu)^2$$

Note, if all claims are litigated individually, then $n = N$, and this expression simplifies to:

$$R_{\text{indiv}} = \sum_{i=1}^{N} (X_i - \mu)^2$$

---

52 Note, in actuality, it is unlikely that $X_i$ has a normal distribution. For example, there is a relatively high probability of a damages value of zero, since this value is associated with the plaintiffs’ failure to prove the defendant’s liability. A mixture distribution may be more appropriate. Regardless, let us assume a normal distribution for purposes of simplicity.
Let us now solve for the optimal sample size—that is, the sample size $n^*$ that minimizes the expectation of $R$, and thus maximizes, on average, the accuracy of the total legal outcome.

The expectation of $R$ can be derived as follows:

$$E(R) = E\left(\sum_{i=1}^{n} (X_i - \mu)^2 n (N - n)(\bar{X}_n - \mu)^2\right)$$

$$= n \sigma^2 n (N - n) \frac{\sigma^2}{n}$$

$$\equiv f(n)$$

Now, to find the minimizing condition, let us solve $\frac{df(n^2)}{dn} = 0$.

$$\min_n f(n) \iff \frac{df(n^2)}{dn} = \frac{d}{dn}\left(n \sigma^2 n (N - n) \frac{\sigma^2}{n}\right) = 0$$

$$\iff n = \sqrt{N}$$

Thus, on average, we minimize $R$, and maximize the accuracy of legal outcomes, by setting $n^* = \sqrt{N}$. Therefore, we maximize accuracy not by litigating each of the claims individually, but rather by sampling and individually litigating $n^* = \sqrt{N}$ claims, and applying the sample mean $\bar{X}_{n^*}$ as the legal outcome in all claims in the extrapolation group.

Let us now relax the assumption of homogeneous claims in order to examine the constraint that claim variability places on the benefits of sampling.

4.2 Heterogeneous Claims

Let us begin by noting that the analysis above allows us to identify bounds on the optimal sampling of claims with respect to claim variability. Specifically, under
circumstances of zero claim variability, our result under the assumption of homogeneous claims applies: we maximize the accuracy of legal outcomes by sampling \( \sqrt{N} \) claims. On the other hand, if claim variability is large, then our optimal sample size is \( N \). Thus, under circumstances of heterogeneous claims, the optimal sample size will fall between \( \sqrt{N} \) and \( N \).

Let us develop a hierarchical model to determine where within these bounds the optimal sample size falls. As before, assume we have a class of \( N \) claims from which we sample \( n \) claims. Now, however, let \( \mu_i \) be the true mean of the verdicts over the total population of possible verdicts in the \( i \)-th claim—i.e., the “correct” verdict in the \( i \)-th claim. Thus, as opposed to circumstances of claim homogeneity, where a single \( \mu \) was used to represent the correct verdict in all \( N \) claims, here we must define a set of terms, \( \mu_1, \mu_2, ..., \mu_N \), to represent the true mean—the “correct” verdict—for each of the \( N \) claims.

Let \( X_i \) be a random variable defined by the actual verdict in the \( i \)-th claim, for \( i = 1, 2, 3, ..., n \), where \( X_i \sim N(\mu_i, \sigma^2) \), independent. Let \( \mu_i \) be a random variable defined as the true mean in the \( i \)-th claim, where \( \mu_i \sim N(\mu, \tau^2) \), i.i.d. And define \( \bar{X}_n \) as the mean of the \( X_i \), for \( i = 1, 2, 3, ..., n \), where \( \bar{X}_n \sim N(\mu, \sigma^2/n \tau^2) \).

Then, pursuant to the above description, let us represent the total measurement error associated with all \( N \) claims with the following loss function:

\[
R = \sum_{i=1}^{n} (X_i - \mu_i)^2 n \sum_{j=n+1}^{N} (\bar{X}_n - \mu_j)^2
\]
It is important to obtain a good estimate of the distribution of the \( \mu_i \)'s. Let us now assume, however, that the \( \mu_i \)'s follow a normal distribution as defined above.

As before, we can obtain the optimal sample size by setting the first derivative of the expectation of the above loss function to zero. We thus find that the optimal sample size under circumstances of heterogeneous claims and judgment variability is

\[
n^* = \sqrt{N} \frac{1}{\sqrt{1 - \lambda^2}} \frac{n \lambda^2}{n^2}
\]

where \( \lambda^2 = \frac{\tau^2}{\sigma^2} \).

This result is intuitive. Consider the following three scenarios. First, if the level of claim heterogeneity is zero (\( \tau^2 = 0 \)), then \( \lambda^2 = 0 \) and \( n^* = \sqrt{N} \), the result obtained from our optimization under circumstances of homogeneous claims above. Second, if \( \tau^2 \geq \sigma^2 \), then \( \lambda^2 \geq 1 \) and \( n^* = N \). Third, if \( 0 < \lambda^2 < 1 \), and \( \tau^2 < \sigma^2 \), then \( n^* = \sqrt{N} \frac{1 + \lambda^2}{1 - \lambda^2} \).

5. Discussion

The current section introduces a number of additional topics for discussion and future research. Specifically, I introduce possibilities for parameter estimation, sequential sampling, stratification and a number of considerations regarding incentives to initiate claims and to settle.

5.1 Parameter Estimation

The current article is intended to introduce a framework for considering accuracy in the law, and in particular, the conditions under which sampling-based aggregation can increase accuracy. Its purpose is not, in general, to make normative statements regarding
whether or how to apply sampling in a given case. However, central to any practical implementation of the framework introduced herein is the ability to estimate the judgment and claim variability parameters. It is beyond the scope of this article to address parameter estimation in detail, but in the current section, I highlight a number of important considerations.

First, it is not necessarily the case that practical application of the framework requires a precise estimation of, or even assignment of numerical values to, the variability parameters. As shown above, when a set of claims is homogeneous (or approximately homogeneous), the optimal sample size is not a function of variability—it is the square root of the number of claims, regardless the value of judgment variability. Further, even in circumstances in which the claims are not approximately homogeneous, the important measurement is the ratio of claim variability to judgment variability—that is, the relative value of claim variability with respect to judgment variability.

Thus, as a first course of action, a court can attempt to identify the sources of claim variability in the class of claims, and can “reduce” such variability through stratification—that is, by breaking the class into multiple subclasses. As in Cimino, discussed above, a court may be able to create multiple relatively homogeneous subclasses rather than a single heterogeneous class. I discuss stratification further below.

At some point of stratification, or perhaps even without stratification, a court may be able to “eyeball” the relative level of claim variability, and avoid the need for formal estimation altogether. In a sense, courts already analyze the variability of claims in the
course of the certification phase of litigation—for example, in their analysis of “commonality.”

Second, if we are interested in finding more accurate estimates of the model’s variability parameters, we have a number of options. It is possible that within a certain population of individuals (e.g., individuals in a given geographic region), jury variability—likely a very significant source of judgment variability—is similar from case to case. In other words, it may not be necessary to study judgment variability within any particular class action lawsuit. Instead, viewing judgment variability as a random variable, it may be possible to understand the distribution of such variability within a given geographic region and category of claims. This is an important area for future research.

We can of course obtain estimates of jury variability, and even of judgment variability, by litigating a single claim multiple times. This can be done within the context of a given class action, or simply in the context of a single claim, or multiple single claims within a given geographic region, in order to study the distribution of judgment variability. As discussed above, replication is costly; but it may be deemed worthwhile in light of the potential benefits of sampling. Further, judgment variability can be estimated using either simulated legal trials or perhaps a set of existing cases.

Estimating claim variability is in some sense more complex than estimating judgment variability. But two possibilities come to mind. First, an area for future research is the possibility of categorizing past legal claims, and undertaking a statistical examination
of claim variability within such categories. This research can be used to formulate claim-variability estimates for future claims.

Second, and perhaps more significantly, courts can address the issue of estimating claim variability through *sequential sampling*, which I address in the following subsection.

### 5.2 Sequential Sampling

The sampling framework introduced above assumes that we designate a sample size prior to beginning the sampling procedure. For example, under conditions of homogeneous claims, we decide upfront that we will include $\sqrt{N}$ claims in the sample group, and $N - \sqrt{N}$ claims in the extrapolation group. Under conditions of heterogeneous claims, we include

$$n^* = \sqrt{N} \frac{1 + \lambda^2}{1 - \lambda^2}$$

in the sample group, and $N - n^*$ claims in the extrapolation group.

A sequential sampling procedure, on the other hand, allows us to begin by taking a small sample, and then to use this sample to estimate relevant parameters prior to determining a more general sample size. In fact, we are not restricted to a two-tier sampling approach. We can repeatedly take small samples, and iteratively update our estimates.

In particular, a court can apply sequential sampling as follows: (i) estimate judgment variability through, for example, one of the procedures discussed above; (ii) estimate a probability distribution for the claim variability parameter and choose a sample size pursuant to the estimate and the sequential sampling plan; (iii) update the estimate pursuant to the claim variability data in the sample, and, if necessary, take a further sample pursuant to the new estimate; and (iv) repeat step three.
By fixing a judgment variability estimate upfront, we can estimate claim variability in the first sample simply by computing the sample variance of the outcomes and adjusting for judgment variability.

The particulars of constructing a sequential sampling scheme are beyond the scope of this article. But note that it is important that the scheme is constructed so as to maintain independence among the claim outcomes, which may involve certain additional costs.

Further, sequential sampling involves some commitment to sampling in the first instance. A court, however, can implement a sequential-sampling procedure in combination with methods discussed in the Parameter Estimation section above. For example, a court may use stratification or simply "eyeballing" to estimate that the class is below a certain level of heterogeneity—and to choose an initial estimate for the claim variability parameter. Then the court can choose a sample size accordingly.

For example, a court may begin in stage one by assuming that a class is approximately homogeneous, and therefore choose a sample group of size \( n^*_\text{homogenous} = \sqrt{N} \), an optimal sample size that does not depend on variability parameters. Once the claims in the sample group have been individually litigated, the court may obtain an estimate of the claim variability.\(^{53}\) Now, since the court has obtained estimates of claim variability and judgment variability, it can relax the assumption of homogeneous claims, and compute

\(^{53}\)Assuming we can obtain a reasonable estimate of judgment variability as discussed above, the court may obtain an estimate of claim variability simply by computing the variability of sample group outcomes and subtracting out the variability attributable to judgment variability.
\[ n^* = \sqrt{N} \frac{1 + \lambda^2}{\sqrt{1 - \lambda^2}}, \] the more-general optimal sample size. If this computation results in a sample size that exceeds the size of the stage-one sample group, then the court will sample an additional \( n^* - n^\text{homogenous} \) claims for litigation. Finally, once the court has litigated \( n^* \) claims and received individualized determinations for these claims, it will assign the mean of these determinations, \( \bar{X}_{n^*} \), to all remaining claims—the claims contained in the extrapolation group.

### 5.3 Stratification

The utility of sampling is associated with the level of homogeneity of the class. At the extremes, a class of perfectly homogeneous claims allows for an optimal sample size of \( n^\text{homogenous} = \sqrt{N} \), while circumstances under which claim heterogeneity is high and judgment variability is low altogether destroy the value of extrapolating from the sample group—that is, such conditions involve an optimal sample size of \( N \). Thus, to simplify the problem and maximize the effectiveness of claim aggregation, a court can stratify a heterogeneous class of claims, thereby creating multiple smaller, but relatively homogeneous, classes.\(^{54}\) Through stratification, the court can “control” for significant sources of variability, and thereby “modify” the functional variability of the claims, and achieve a level of efficiency and accuracy that is otherwise not possible.

As a simple example, assume that 5,000 smoker claims can be distinguished from one another only on the basis of the type of damages suffered. One-fifth of the plaintiffs

---

\(^{54}\) See generally Cimino, 751 F. Supp. at 649; Saks & Blanck, supra note 2, at 842.
suffered from heart disease, one-fifth suffered from lung disease, one-fifth suffered merely from over paying, etc. Assume, for simplicity, no overlap among the categories. Rather than sampling from the entire class of 5,000 claims \( n^* = \sqrt{N} \left(\frac{1 + \lambda^2}{1 - \lambda^2}\right) \) claims—a relatively large sample size, if not the size of the entire 5,000-claim class altogether—the court can create five 1,000-claim homogeneous classes, from which it will sample \( n_i^* = \sqrt{N} \) claims from each (where \( i \) is an index for subclasses 1 ... 5). Thus, in total, the court will sample \( 5(\sqrt{N}^2 \) claims; but, as in Cimino, it will use the \( n_1^* = \sqrt{N_1} \) claims to extrapolate determinations for the first subclass, the \( n_2^* = \sqrt{N_2} \) claims to extrapolate determinations for the second subclass, and so on.

An interesting topic for future research is the problem of optimal stratification—\( i.e., \) the number of strata—under various variability conditions.

5.4 Incentives to Initiate Claims and to Settle

As mentioned above, it is not my aim in the current article to make normative determinations with respect to the implementation of claim-aggregation procedures. However, an important consideration for lawmakers is the way in which such a scheme might affect incentives to initiate actions, as well as incentives to settle claims (rather than litigate), given that an action has been initiated.

At first glance, one might think that claim aggregation may have a negative effect, with respect to social objectives, on incentives of potential litigants to initiate claims.

Specifically, within a given class of claims, potential litigants with weaker claims (\( i.e., \) claims involving a lower expected recovery) may be overincentivized to join a class governed by a
procedure involving aggregation of claims, while potential litigants with stronger claims may be underincentivized to join such a class. This is because claim aggregation involves a substantial possibility that a given litigant will be chosen for the extrapolation group, in which case, she will receive the mean determination in the sample group rather than an individualized determination.

There are numerous problems with this reasoning; and to the extent that incentives are necessarily affected, courts may be able to address such effects at a relatively low cost. First, to the extent that “weaker” plaintiffs are overincentivized to join a class under the aggregation scheme, it is not dissimilar to the overincentive produced by the current class action framework, wherein a verdict associated with “representative” claims is applied to the entire class. Of course, a court may order a particular procedure, such as individualized hearings, for determining the allocation of the total award among the claimants; but such a procedure is entirely consistent with the aggregation scheme as well. For example, a court may employ sampling for the purpose of arriving at a classwide liability or damages determination, but thereafter may impose a further procedure for determining the allocation of the damages award.

Under both the current class action framework and a sampling framework, courts may avoid undesirable incentives by determining that a class is too heterogeneous, thus refusing to certify the class for class action litigation. Further, a court may seek to stratify the class in the first instance and thereby create relatively homogeneous subclasses. In this way, the weaker litigants are not overincentivized, and the stronger litigants are not
underincentivized, to join the class. Each category of litigants would expect to be grouped with other similarly-situated litigants.

Additionally, it is well recognized that sampling offers the potential for enormous cost savings, and to the extent that aggregation allows for increased accuracy and reductions in litigation costs, incentives to litigate claims thereby become more aligned with social objectives.\(^{55}\) Assertions that sampling increases social costs by burdening the court system with numerous bellwether trials are flawed. Such assertions are based on the assumption that, if a court refuses to certify a proposed class, the members of the class will treat the ruling as a general loss and will not bring their claims individually. To the extent that this is in fact the case (it often is), closing the court’s doors on litigants with valid claims certainly should not be understood as a social victory. To the contrary, it represents a failure of the judicial system, and the social costs of such failures—including, for example, the harmful incentive structure established by allowing mass tortfeasors to avoid liability—are likely to far outweigh the litigation costs of sampling.\(^{56}\)

---

\(^{55}\)This applies whether social objectives relate, for example, to concerns of deterrence or concerns of fairness.

\(^{56}\)See generally Blue Cross and Blue Shield of New Jersey, Inc., 178 F. Supp. 2d at 247-48 (“In mass exposure cases with hundreds of thousands or millions of injured ... the cost of such one-on-one procedures is insuperable and unsuitable for either a jury or a bench trial. The consequence of requiring individual proof from each smoker would be to allow defendants who have injured millions of people and caused billions of dollars in damages, to escape all liability.”); David Rosenberg, The Causal Connection in Mass Exposure Cases: A “Public Law” Vision of the Tort System, 97 Harv. L. Rev. 849 (1984); David Rosenberg, Individual Justice and Collectivizing Risk-Based Claims in Mass-Exposure Cases, 71 N.Y.U. L. Rev. 210 (1996). Note, to the extent that it may be argued that sampling procedures allow for increased accuracy and should therefore occur even when such procedures are not required to maintain a class action, sampling arguably results in heightened litigation costs—e.g., from litigating \(\sqrt{N}\) claims rather than a single claim—and will therefore affect incentives to litigate and to settle. Claimants may be less willing to bring claims, and claimants and defendants more eager to settle matters out of court, to avoid the relatively high costs of litigation. It may be argued, however, that the costs of judgment variability, and the inaccuracy that results therefrom, outweigh any increase in litigation cost, including altered incentives with respect to litigation and settlement. That is, current class action procedures, where a single trial—subject to a high degree of judgment variability—can possibly determine wholesale the outcome of
Further, aggregation procedures may be designed to encourage settlement among the parties. Indeed, the use of bellwether trials for purposes of facilitating settlement and the resolution of claims is well recognized. As the Fifth Circuit remarked in *In re Chevron U.S.A., Inc.*, “The notion that the trial of some members of a large group of claimants may provide a basis for enhancing prospects of settlement or for resolving common issues or claims is a sound one that has achieved general acceptance by both bench and bar.”

For example, a court may opt for a framework in which sampled claims are litigated in stages. This not only allows for calibration of sampling procedures, as described above, but also allows the parties to calibrate their expectations with respect to the outcome of the litigation, thus facilitating settlement. The Fifth Circuit explained, “The reasons for acceptance by bench and bar are apparent. If a representative group of claimants are tried to verdict, the results of such trials can be beneficial for litigants who desire to settle such claims by providing information on the value of the cases as reflected by the jury verdicts.”

Thousands, or even millions, of claims arguably entail far greater costs than those associated with litigating the claims contained in the sample group. Future research should examine ways in which courts can take advantage of economies of scale to reduce litigation costs associated with sampling. Additionally, future research should investigate the effects on incentives to initiate claims and settle suits after accounting for the role of law firms and contingency fees in particular. It is possible that such effects are mitigated after accounting for firms’ willingness to incur the risk associated with lawsuits in exchange for larger potential payoffs. It is also possible that the costs associated with aggregation combined with the resulting increases in accuracy may in fact result in a reduction in frivolous or otherwise undesirable litigation. In any event, it should not be assumed that a claimant in the sample group would incur costs beyond those incurred by claimants in the extrapolation group; such costs can be spread across the class, if the class (rather than, e.g., a law firm) is incurring the costs of litigation in the first instance.


58 *In re Chevron U.S.A., Inc.*, 109 F.3d at 1019; *see also* Connecticut Cooling Total Air, Inc. v. Connecticut Natural Gas Corp., 46 Conn. Supp. 82, 89 (Conn. Super. Ct. 1999) (“The court has devised a case management strategy
A defendant corporation that observes consistent losses in the first stage of sampling may decide to bend to the plaintiffs' settlement demands. Similarly, a plaintiff class that observes consistent losses may decide to drop further action altogether. Between these two extremes, the parties can calibrate their expectations and their demands, and thereby enhance the likelihood of settling after perhaps the initial stage of litigation.

6. Conclusion

Courts and authors have suggested that, under certain circumstances, sampling procedures can not only increase efficiency, but accuracy as well. Saks and Blanck, as well as other authors, have highlighted the role of what I have referred to as “judgement variability” in the law, and the ability to employ sampling procedures to reduce such variability and thereby reduce error in the law. Such assertions have been used to rebut anti-aggregation arguments that are based on the premise that accuracy cannot be sacrificed for the sake of efficiency.

But under what circumstances in particular can sampling procedures increase accuracy? What are the conditions under which sampling will result in a net reduction—accounting for effects of both judgment variability and claim variability—of error in the law? Assertions that sampling procedures can increase accuracy have been met with skepticism and a general unwillingness to rely on such assertions in real-world contexts. Such skepticism is arguably due to the fact that legal scholarship has not begun to

---

involving trial of a 'bellwether' case from among these cases, as a potential means of achieving a more prompt trial of issues that may lead to negotiated resolutions of all the cases.’). See generally Cimino, 751 F. Supp. at 649.
contemplate, in any rigorous form, the practical effect of sampling procedures on accuracy under real-world conditions of both claim and judgment variability, and with realistic constraints imposed by the law.

In the current article, I have introduced a framework for examining the conditions under which sampling can increase accuracy in the law. In particular, I have introduced a simple model for studying the effects of sampling on accuracy, and for deriving the optimal sample size, with respect to accuracy, under conditions of claim and judgment variability, and with constraints described by reductive sampling. I have also introduced a number of important considerations and topics for future research. Specifically, I have discussed possibilities for estimating variability parameters and the use of sequential sampling, the potential for stratification procedures that may allow courts to reduce the degree of claim heterogeneity and thereby increase the utility of claim aggregation, and considerations related to incentives to initiate claims and to settle under circumstances of claim aggregation.

My aim in the current article was to build on previous literature to introduce a framework and discussion for considering the conditions under which sampling can increase accuracy in the law. My hope is that the article will serve as a first step toward a more rigorous examination of such conditions, and ultimately the practical implementation of sampling to increase accuracy, as well as efficiency, in the law.
II. Sampling and Reliability in Class Action Litigation*

Courts continue to struggle with the limits of statistical sampling in resolving claims arising from “the mass repetitive wrong.” In *Cimino v. Raymark Indus., Inc.*, a 1990 asbestos class action, U.S. District Court Judge Robert Parker divided a class of 2,298 claimants into five subclasses based on claimed injuries, and selected a random sample from each subclass to form a representative sample group of 160 claimants. He submitted the sample group’s claims to a jury for individual determinations and then extrapolated outcomes for the non-sampled claims by applying the average sample group determinations to each subclass, respectively. The Fifth Circuit ultimately rejected Judge Parker’s sampling procedure on constitutional grounds.

In recent years, courts and commentators have criticized the use of sampling to prove classwide liability and damages. It is widely believed that sampling serves goals of

---

* This article originally appeared in 2016 CARL. REV. DE NOVO 207 (2016). Citation: Hillel J. Bavli, *Sampling and Reliability in Class Action Litigation*, 2016 CARL. REV. DE NOVO 207 (2016). It may have been modified from its original form for purposes of consistency. The author thanks John Kenneth Felter and the Cardozo Law Review Editorial Board for their helpful comments and edits.

1 *Cimino v. Raymark Indus., Inc.*, 751 F. Supp. 649, 652 (E.D. Tex. 1990) (citing HERBERT B. NEWBERG, NEWBERG ON CLASS ACTIONS § 17.06, at 373 (2d ed. 1985)).

2 Id. at 652–54; see also Hillel J. Bavli, *Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation*, 14 LAW, PROBABILITY & RISK 67, 68 (2015) [hereinafter *Aggregating for Accuracy*].

3 *Aggregating for Accuracy*, supra note 2, at 68.

4 See *Cimino v. Raymark Indus., Inc.*, 151 F.3d 297 (5th Cir. 1998).

efficiency, but only at the cost of reliability—and that the “benefits of efficiency can never be purchased at the cost of fairness.”

In a 1992 Stanford Law Review article, Professors Michael Saks and Peter Blanck argued that sampling, when performed correctly, not only satisfies the standards of reliability achievable through individual litigation, but that sampling can also increase the reliability of legal outcomes.

However, courts have generally been unwilling to accept the reliability of sampling procedures used to prove classwide liability and damages; and courts frequently reject sampling on constitutional and procedural grounds. But, as Saks and Blanck have asserted, “a major—perhaps the major—due process concern in an aggregated trial is the validity of the outcome.” Moreover, “[t]he main argument against trial by aggregation and sampling asserts that such trials cannot give the parties as accurate a result as they would obtain through traditional bilateral trials.”

---


10 Saks & Blanck, supra note 7, at 833; see also In re Chevron U.S.A., Inc., 109 F.3d at 1020 (“[O]ur procedural due process concerns focus on the fact that the procedure embodied in the district court’s trial plan is devoid of safeguards designed to ensure that the claims against Chevron of the non-represented plaintiffs as they relate to liability or causation are determined in a proceeding that is reasonably calculated to reflect the results that would be obtained if those claims were actually tried.”).
The question remains: why have courts and commentators largely discounted the argument that sampling offers not only increased efficiency, but reliability as well? In a 2015 paper, I explain that a primary source of skepticism may be the generality with which the argument has been made.\textsuperscript{11} Although Saks and Blanck, and others, have suggested that sampling may increase the reliability of legal outcomes, the literature has not adequately developed the argument or produced a framework to analyze the effect of sampling on reliability.\textsuperscript{12} Indeed, sampling does not inevitably increase the reliability of legal outcomes; rather, its effect on reliability depends on the particular features of the claims.\textsuperscript{13} Therefore, in \textit{Aggregating for Accuracy}, I develop a formal framework for examining conditions under which sampling can increase the reliability of legal outcomes.\textsuperscript{14}

In this Article, I explain my conclusions in \textit{Aggregating for Accuracy} in non-mathematical terms, and underscore certain implications with respect to class action litigation and considerations in light of the U.S. Supreme Court’s recent decision in \textit{Tyson Foods, Inc. v. Bouaphakeo}.\textsuperscript{15} I begin by describing the building blocks of my analysis—the concepts of reliability and accuracy in the law.

1. \textbf{Reliability in the Law}

---

\textsuperscript{11} \textit{Aggregating for Accuracy}, supra note 2, at 69.

\textsuperscript{12} Id.

\textsuperscript{13} Id.

\textsuperscript{14} See generally id.

\textsuperscript{15} 136 S. Ct. 1036 (2016).
Assume that for every legal claim there is a “correct” outcome that can be determined by applying the “true” state of the law to the complete facts surrounding the claim. But, in the real world, complete knowledge regarding the law and the facts surrounding a claim is unavailable. Consequently, a legal outcome (resulting from a trial or other adversarial proceeding) serves as an *estimate* of the “correct” outcome for the claim. Therefore, for every legal claim, there is an error term that (although generally unknown) reflects the disparity between the observed outcome and the “correct” outcome. *Accuracy*, then, is defined in terms of the proximity of the observed outcome to the “correct” outcome.

One convenient and arguably sensible conceptualization of the “correct” outcome associated with a legal claim is the award that would result from computing the mean of infinitely many adjudications of the claim under various conditions (e.g., various jury compositions, judges, attorneys, presentations of evidence, etc.).

Thus, I define the *reliability* of a legal procedure as the accuracy of the legal outcome that can be expected by following the procedure. For example, if it is expected that a certain

---

16 Alternatively, it can be assumed that there is a distribution of “correct” outcomes associated with every legal claim. See generally Aggregating for Accuracy, supra note 2, at 69.

17 See generally id. at 74; Jonathan J. Koehler & Daniel Shaviro, Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods, 75 CORNELL L. REV. 247 (1990) (discussing “accuracy”). Note that accuracy can be defined more formally using the concepts of bias and variance. However, an explicit discussion of these concepts is beyond the scope of this Article.

18 See Saks & Blanck, supra note 7, at 833–34; Aggregating for Accuracy, supra note 2, at 74. Other measures of central tendency, such as the median, are also possible. For simplicity, throughout this Article we use solely the mean.
legal procedure will produce a highly accurate outcome—an observed outcome in close proximity to the “correct” outcome—the procedure is considered reliable.\(^{19}\)

An in-depth discussion of the role of accuracy in the law is beyond the scope of this Article. I assume that accuracy is a fundamental goal of the law, whether such goal is grounded in accuracy itself or a further aim, such as deterrence or fairness.\(^{20}\)

2. **Claim and Judgment Variability**

Critics of sampling have focused on the error that results from applying a point estimate—e.g., the average of the sample group outcomes—to a class of heterogeneous legal claims.\(^{21}\) Heterogeneity can be described in terms of *claim variability*—i.e., the variability of the facts of the claims, or, more precisely, the variability of the “correct” outcomes associated with individual claims.\(^{22}\)

But critics of sampling have ignored a second type of error: error resulting from *judgment variability*.\(^{23}\) Judgment variability represents the randomness associated with a claim’s outcome. It arises from the variability in conditions under which outcomes are determined, including the composition of the jury, the judge presiding over the case, the

\(^{19}\) See *Aggregating for Accuracy*, supra note 2, at 74–78 for a detailed discussion of accuracy and variability in the law.

\(^{20}\) See *id.* at 75; Saks & Blanck, supra note 7, at 829 (citing Redish & Marshall, *supra* note 9, at 476–77).

\(^{21}\) *Aggregating for Accuracy*, supra note 2, at 75–76. See generally Saby Ghoshray, *Hijacked by Statistics, Rescued by Wal-Mart v. Dukes: Probing Commonality and Due Process Concerns in Modern Class Action Litigation*, 44 LOY. U. CHI. L.J. 467 (2012). If there are subclasses, then a court is likely to assign the average sample group outcome within each subclass to the non-sampled claims in each subclass, respectively. *See, e.g.*, Cimino v. Raymark Indus., Inc., 751 F. Supp. 649, 651–54 (E.D. Tex. 1990).

\(^{22}\) *Aggregating for Accuracy*, supra note 2, at 76.

\(^{23}\) *Id.* at 76–78.
presentation of evidence, the attorneys involved in the case, etc.24 For example, if a single claim is tried ten times independently (each trial with a new selection of trier of fact, attorneys, etc.), it is likely that there would be ten distinct verdicts. But, if there is a single correct outcome associated with the claim, then judgment variability reflects error—disparities between the observed outcomes and the “correct” outcome.25

For simplicity, I assume that, on average, a case will result in the “correct” outcome—in statistical terms, the outcome is unbiased. Judgment variability represents the degree to which an observed outcome varies around the “correct” outcome.26 The reader may imagine a bell curve centered at the “correct” outcome, where the judgment variability determines the width of the curve. On average, the outcome will be the “correct” outcome; but the outcome is variable around the “correct” outcome, and such randomness, represented by judgment variability, reflects error.27

In determining appropriate standards in the class action context, courts and commentators have focused on error resulting from claim variability, but have generally ignored error resulting from judgment variability.28 Indeed, even the purported ideal of individual litigation produces outcomes that are subject to judgment variability.

24 Id. at 76–77.


26 See generally Aggregating for Accuracy, supra note 2, at 75–78.

27 See id. at 75–79. Consider the magnitude of error that results from judgment variability when classwide damages are determined by adjudicating a single representative claim.

28 Id. at 76–78.
3. **Reducing Judgment Variability with Sampling**

Consider a costly procedure through which the outcome of a claim is determined by averaging the verdicts resulting from ten independent “replications” of a trial, or “repeated adjudications” (involving, for example, different judges, juries, attorneys, presentations of evidence, etc.). Assuming the outcome is relatively unbiased, it is easy to show that following this procedure results in an accurate outcome—an outcome that is close to the “correct” outcome. Similarly, this procedure will produce an accurate outcome for each claim of each member of a putative class (or subclass). Replication thus increases the reliability of legal outcomes by reducing the error caused by judgment variability.

On the other hand, the costs of such a procedure are enormous, and likely not justified by the benefits of the procedure for a single claim (or for each putative class member’s claim that requires application of the procedure for the individual claim).

Importantly, however, in a class of homogeneous claims, it is not necessary to follow the procedure for each class member’s claim to realize the benefits of the procedure: following the replication procedure for a single claim and extrapolating the result to all claims of the homogeneous class (or subclass) will yield the reliability benefits as though the procedure was followed for each claim.

Class actions provide opportunities to realize the reliability benefits of replication without incurring the prohibitive costs that such procedures entail.\(^29\) In particular, courts can use sampling to improve legal outcomes by reducing error resulting from judgment

\(^{29}\) *Id.* at 77.
variability. Moreover, sampling may offer a degree of reliability that cannot be obtained even through the purported ideal of individual adjudication. Further, although homogeneity is helpful, it is not necessary. As explained below, the benefits of sampling in a heterogeneous class depend on whether, and to what extent, the error resulting from judgment variability dominates the error resulting from claim variability.\textsuperscript{30}

4. **Conditions Under Which Sampling Can Improve the Reliability of Legal Outcomes**

Above, I describe how repeated adjudication can increase the reliability of legal outcomes. Assuming a certain degree of homogeneity, if a court were unconstrained by cost or law in its pursuit of reliability, it might “sample” all the claims of a class for individual adjudication and then replace all individual outcomes with a single aggregated outcome. In fact, a court could further reduce the error resulting from judgment variability by adjudicating each claim twice, or more times for that matter. But, in addition to the high costs of litigation, a court’s ability to achieve extremely reliable outcomes is constrained by law.

In *Aggregating for Accuracy*, I argue that a court may not replace an individualized outcome with an aggregated one.\textsuperscript{31} I highlight a fundamental distinction, based on constitutional law and rules of civil procedure, between a court’s authority to extrapolate a representative claim’s determination to unadjudicated claims and its authority to replace

\textsuperscript{30} *Id.* at 78; see Saks & Blanck, *supra* note 7, at 833–37.

\textsuperscript{31} *Aggregating for Accuracy, supra* note 2, at 78–80.
an individually adjudicated outcome with an aggregated one.\textsuperscript{32} A detailed discussion of this issue is beyond the scope of this Article. Assume, therefore, that a court must choose between sampling a claim for individual adjudication and preserving the claim for assignment of an aggregated outcome extrapolated from the sampled claims.\textsuperscript{33}

This constraint can be modeled using a type of exploration-exploitation tradeoff I call \textit{reductive sampling}, in which sampling a unit—here, a legal claim—means reducing (or, as here, eliminating) its eligibility for later extrapolation with respect to that unit.\textsuperscript{34}

In \textit{Aggregating for Accuracy}, I show that, using the reductive sampling framework and standard statistical tools for minimizing error, in the context of a class of homogeneous (or relatively homogeneous) claims, accuracy is maximized, not by adjudicating each of the claims individually (as is often viewed as ideal), but rather by determining individual outcomes for a small random sample of claims, and assigning the average of the sample group outcomes to the remaining, non-sampled, claims.\textsuperscript{35} In particular, I show that, for a class of $N$ homogeneous claims, a court maximizes accuracy by sampling $\sqrt{N}$ claims, rather than all $N$ claims, for individual adjudication.\textsuperscript{36} For example, if the class contains 5,000 homogeneous claims, a court can maximize accuracy by randomly sampling about 70

\begin{thebibliography}{9}
\bibitem{note32} Id.
\bibitem{note33} See id. at 80–81.
\bibitem{note35} See \textit{Aggregating for Accuracy}, supra note 2, at 81–82.
\bibitem{note36} Id.
\end{thebibliography}
claims for individual adjudication, assigning the 70 individual outcomes to the sampled claims, and assigning the arithmetic mean of the 70 sample outcomes to the remaining 4,930 claims.

\( \sqrt{N} \) (or about 70 in the example above) is the number that balances, on the one hand, a court’s interest in obtaining information regarding the “correct” outcome of the homogeneous claims (which increases with sample size), and, on the other hand, its interest in preserving claims to which to assign the accurate aggregated outcome.\(^\text{37}\) The 70 claims in the sample group receive individual adjudications that are subject to significant error caused by judgment variability, whereas the remaining 4,930 claims in the extrapolation group are assigned aggregated outcomes that have been “refined” by repeated adjudication—i.e., outcomes whose judgment variability has been reduced significantly by averaging over approximately 70 repeated adjudications.\(^\text{38}\)

Now, to examine sampling in the context of heterogeneous claims, consider an additional factor: although the sampling procedure described above reduces error caused by judgment variability, assigning a single aggregated outcome (e.g., the average of the sample group adjudications) as the estimate of the “correct” outcomes associated with a group of heterogeneous claims—claims that actually involve numerous distinct “correct” outcomes—introduces error reflecting the disparities between the estimate and each of the “correct” outcomes.\(^\text{39}\) Thus, as the heterogeneity of the class increases, the value of

---

\(^\text{37}\) See id. at 80–81.

\(^\text{38}\) Id. at 80–82.

\(^\text{39}\) Id. at 82–83.
assigning a single aggregated outcome as the estimate of the “correct” outcomes
decreases.\textsuperscript{40} Further, at some degree of heterogeneity, the benefits of sampling, with
respect to judgment variability, are outweighed by the detriments of sampling, with respect
to claim variability.\textsuperscript{41}

In \textit{Aggregating for Accuracy}, I show that if claim variability is zero, then the optimal
sample size is $\sqrt{N}$, the homogeneous optimum; if claim variability is greater than judgment
variability (e.g., in terms of damages awarded), the optimal sample size is $N$, which is
equivalent to individual adjudications; and finally, if judgment variability is greater than
claim variability, then the optimal sample size is between $\sqrt{N}$ and $N$, and can be determined
by a particular formula (derived in \textit{Aggregating for Accuracy}) involving the number of
claims in the class, claim variability, and judgment variability.\textsuperscript{42}

5. Implications and Conclusions

The conclusions explained above, and derived in \textit{Aggregating for Accuracy}, have
important implications for a court’s treatment of statistical sampling in class action
litigation. Perhaps most importantly, a court should not assume, as many courts have, that
sampling reduces reliability. The discussion above makes clear that sampling may enhance
reliability as well as efficiency.

For example, these conclusions have important implications in the context of class
actions brought under Rule 23(b)(3) of the Federal Rules of Civil Procedure. In particular,

\begin{itemize}
\item \textsuperscript{40} Id.
\item \textsuperscript{41} See id. at 82–83. See generally Saks & Blanck, \textit{supra} note 7.
\item \textsuperscript{42} See \textit{Aggregating for Accuracy}, \textit{supra} note 2, at 83.
\end{itemize}
class representatives have proposed sampling procedures for purposes of fulfilling the Rule's requirement that “the questions of law or fact common to class members predominate over any questions affecting only individual members.” In addressing this requirement, the class representatives have offered sampling-based methodologies at the class certification stage in attempt to demonstrate that liability and damages can be determined on a classwide basis. Courts regularly reject such attempts, however, on grounds—implicitly or explicitly—of reliability. But, for the reasons discussed above, courts are generally not justified in assuming that sampling diminishes reliability.

The Supreme Court, in its recent decision in Tyson Foods, Inc. v. Bouaphakeo, refused to adopt “a broad rule against the use in class actions of . . . representative evidence”—specifically, evidence based on “a representative or statistical sample” offered to establish classwide liability. Instead, it held that “[w]hether and when statistical evidence can be used to establish classwide liability will depend on the purpose for which the evidence is being introduced and on ‘the elements of the underlying cause of action.’”

Tyson Foods involved claims by Tyson employees, certified by the district court as a Rule 23(b)(3) class, alleging that their employer, Tyson, failed to pay compensable overtime wages under the Fair Labor Standards Act (FLSA) and Iowa law for time spent

---

43 See Comcast Corp. v. Behrend, 133 S. Ct. 1426, 1432 (2013) (quoting FED. R. CIV. P. 23(b)(3)).


45 Tyson Foods, Inc., 136 S. Ct. at 1046.

46 Id. (quoting Erica P. John Fund, Inc. v. Halliburton Co., 563 U.S. 804, 809 (2011)).
“donning and doffing protective gear.” The district court submitted the issues of liability and damages to the jury, which “returned a special verdict finding that time spent in donning and doffing protective gear . . . was compensable work,” and “awarded the class about $2.9 million in unpaid wages.” Tyson appealed, arguing, inter alia: “[T]he class should not have been certified because the primary method of proving injury assumed each employee spent the same time donning and doffing protective gear, even though differences in the composition of that gear may have meant that, in fact, employees took different amounts of time to don and doff.” The Eighth Circuit Court of Appeals affirmed the judgment and the award.

In ruling that the district court did not err in certifying the class, the Supreme Court held:

In many cases, a representative sample is “the only practicable means to collect and present relevant data” establishing a defendant’s liability. Manual of Complex Litigation §11.493, p. 102 (4th ed. 2004). In a case where representative evidence is relevant in proving a plaintiff’s individual claim, that evidence cannot be deemed improper merely because the claim is brought on behalf of a class. To so hold would ignore the Rules Enabling Act’s pellucid instruction that use of the class device cannot “abridge . . . any substantive right.” 28 U.S.C. § 2072(b).

The Court continued: "One way for respondents to show, then, that the sample relied upon here is a permissible method of proving classwide liability is by showing that

---

47 Tyson Foods, Inc., 136 S. Ct. at 1042.

48 Id. at 1044. Note that a class expert recommended an award of approximately $6.7 million. Id. at 1052.

49 Id. at 1041.

50 Id. at 1044; Bouaphakeo v. Tyson Foods, Inc., 765 F.3d 791 (8th Cir. 2014).

51 Tyson Foods, Inc., 136 S. Ct. at 1046.
each class member could have relied on that sample to establish liability if he or she had brought an individual action.”\textsuperscript{52} It held that “[i]f the sample could have sustained a reasonable jury finding as to hours worked in each employee's individual action, that sample is a permissible means of establishing the employees' hours worked in a class action.”\textsuperscript{53}

The Court clarified its ruling in \textit{Wal-Mart Stores, Inc. v. Dukes},\textsuperscript{54} explaining that “\textit{Wal-Mart} does not stand for the broad proposition that a representative sample is an impermissible means of establishing classwide liability.”\textsuperscript{55} \textit{Wal-Mart} involved a class of approximately 1.5 million current and former female employees alleging gender discrimination under Title VII.\textsuperscript{56} The Court explained that “[t]he plaintiffs in \textit{Wal-Mart} did not provide significant proof of a common policy of discrimination to which each employee was subject.”\textsuperscript{57} Ultimately, the Supreme Court rejected plaintiffs’ proposed methodology—proposed “as a means of overcoming the absence of a common policy”\textsuperscript{58}—by which a sample of class members would be selected for individual determinations of liability and backpay, and the aggregated damages award would be derived by extrapolating the

\textsuperscript{52} \textit{Id.}

\textsuperscript{53} \textit{Id. at} 1046–47. The Court explained: “If the employees had proceeded with 3,344 individual lawsuits, each employee likely would have had to introduce [the class expert’s] study to prove the hours he or she worked. Rather than absolving the employees from proving individual injury, the representative evidence here was a permissible means of making that very showing.” \textit{Id.} at 1047.

\textsuperscript{54} 564 U.S. 338 (2011).

\textsuperscript{55} \textit{Tyson Foods, Inc.}, 136 S. Ct. at 1048.

\textsuperscript{56} \textit{Id.}

\textsuperscript{57} \textit{Id.}

\textsuperscript{58} \textit{Id.}
“number of (presumptively) valid claims” from the percentage of the sampled claims determined to be valid, and then multiplying this number by “the average backpay award in the sample set.”59

The Court explained that its holding in Tyson Foods is “in accord with” its decision in Wal-Mart: “Since the Court held that the employees were not similarly situated, none of them could have prevailed in an individual suit by relying on depositions detailing the ways in which other employees were discriminated against by their particular store managers.”60 The Court explained that “[b]y extension, if the employees had brought 1 ½ million individual suits, there would be little or no role for representative evidence,” and that “[p]ermitting the use of that sample in a class action, therefore, would have violated the Rules Enabling Act by giving plaintiffs and defendants different rights in a class proceeding than they could have asserted in an individual action.”61 In Tyson Foods, the Court held:

[T]he study here could have been sufficient to sustain a jury finding as to hours worked if it were introduced in each employee’s individual action. While the experiences of the employees in Wal-Mart bore little relationship to one another, in this case each employee worked in the same facility, did similar work, and was paid under the same policy.62

59 Id. (quoting Wal-Mart Stores, Inc. v. Dukes, 564 U.S. 338, 367 (2011)).

60 Tyson Foods, Inc., 136 S. Ct. at 1048.

61 Id.

62 Id. at 1048. The Supreme Court held that, although the “question whether uninjured class members may recover is one of great importance,” it is not “a question yet fairly presented by this case, because the damages award has not yet been disbursed, nor does the record indicate how it will be disbursed.” Id. at 1050. The Court remanded the case for further proceedings consistent with its opinion. Id. Significantly, Chief Justice Roberts expressed concern, in a concurring opinion joined in part by Justice Alito, that, since the district court may be unable to “fashion a method for awarding damages only to those class members who suffered an actual injury,” id. at 1050, “it remains to be seen whether the jury verdict can stand.” Id. at 1053. Additionally, Justice Thomas, in a dissenting opinion joined by Justice Alito, asserted that “[t]he District Court erred at the class certification stage by holding that the plaintiffs satisfied Rule 23’s predominance
Thus, although the Supreme Court explicitly refused to adopt “broad and categorical rules governing the use of representative and statistical evidence in class actions,” the *Tyson Foods* decision approves the use of representative evidence to establish classwide liability.63

Distinguish two types of sampling methodologies that putative class representatives have attempted to use for establishing classwide liability and damages: 1) the use of “representative evidence” offered to the trier of fact as probative of classwide liability64 or damages, and 2) the use of representative adjudications to extrapolate outcomes for non-sampled (i.e., non-adjudicated) claims.65 Although the Supreme Court did not distinguish between these two types of sampling methodologies, it is likely that *Tyson Foods* expands the ability of class representatives to use *representative evidence* to establish classwide liability in particular.

Regarding the use of representative evidence, the Court held: “A representative or statistical sample, like all evidence, is a means to establish or defend against liability. Its permissibility turns not on the form a proceeding takes—be it a class or individual action—but on the degree to which the evidence is *reliable* in proving or disproving the elements of

---

63 *Id.* at 1049.

64 *See id.* at 1043.

65 *See* Aggregating for Accuracy, *supra* note 2, at 70–72 (citing cases and literature); *see, e.g.*, Cimino v. Raymark Indus., Inc., 751 F. Supp. 649, 652–54 (E.D. Tex. 1990).
the relevant cause of action." The Court's decision in *Tyson Foods* approves the use of statistical sampling to establish classwide liability; but the role of *reliability* in determining whether a court permits statistical sampling to prove classwide liability remains central to the analysis.

The conclusions explained above suggest that, while a court should examine the reliability of statistical sampling for, among other things, methodological flaws and issues related to the cohesiveness of the class, it should not discount the reliability of statistical sampling—and representative evidence in particular—because of the sampling itself. Indeed, as explained above, sampling may improve reliability as well as efficiency.

Additionally, it is important to realize that, although the discussion above relates particularly to the reliability benefits of representative adjudications, the conclusions generally apply to representative evidence as well. As explained, repeated adjudication may improve reliability by enabling a court, in essence, to incorporate additional information regarding a class of claims—e.g., by averaging over multiple adjudications rather than relying on a single adjudication—and thus minimize error caused by judgment variability. Repeated adjudication in a heterogeneous class similarly confers reliability benefits, as long as the error-reducing benefits of "information sharing" with respect to judgment variability outweigh the error-inducing costs with respect to claim variability.

The use of representative evidence similarly improves reliability. Although the method of "information sharing" using representative evidence is different—involving, for

---


65
example, the additional step of providing the information to the trier of fact, rather than incorporating it in the outcome directly (e.g., by averaging over repeated adjudications)—insofar as the trier of fact incorporates the representative evidence in its determination of the outcome, the reliability benefits of repeated adjudication apply similarly to the use of representative evidence.67

Tyson Foods is likely to have a significant impact on class action litigation. Putative class representatives will be encouraged to use representative evidence to establish classwide liability, and perhaps damages as well. Using arguments establishing, for example, that “each class member could have relied on th[e] sample to establish liability if

---

67 In particular, in Aggregating for Accuracy, an award in a heterogeneous class is modeled hierarchically: the “correct” awards in the class are distributed around some global mean, whereas each actual award is “drawn” from a distribution around each claim’s “correct” award. See generally Aggregating for Accuracy, supra note 2, at 82–83. The former distribution represents claim variability, whereas the latter distribution represents judgment variability. In a homogeneous class, replication offers accuracy benefits by providing additional information regarding the “correct” award associated with the replicated claim, which otherwise would be obscured by judgment variability. In a heterogeneous class, sampling offers accuracy benefits, with respect to a certain claim, not by providing information regarding that claim’s “correct” award directly, but by providing information regarding the global mean around which all of the “correct” awards are distributed, and thereby regarding the “correct” award for the subject claim indirectly. Similarly, representative evidence, such as the type in dispute in Tyson Foods (where, for example, the sample reflects variability of measured donning and doffing times rather than judgment variability), offers accuracy benefits, with respect to a certain claim, by providing information regarding the global mean around which the “correct” awards are distributed, and thereby regarding the “correct” award for that claim in particular. Another way of understanding this is through “comparable-case guidance” (CCG) methods, whereby a court uses information regarding awards in prior comparable cases as guidance for a fact-finder’s determination of damages. See Hillel J. Bavli, The Logic of Comparable-Case Guidance in the Determination of Awards for Pain and Suffering and Punitive Damages, U. CIN. L. REV. (forthcoming 2017) [hereinafter The Logic of CCG]. The Logic of CCG examines the statistical mechanism by which CCG affects awards, and the conditions under which such evidence will improve accuracy. In particular, the paper explains that, under certain mild behavioral assumptions, the risk that such evidence would reduce accuracy—that error resulting from claim variability and bias would outweigh the accuracy benefits of reducing judgment variability—is minimal. See id. Like CCG, representative evidence provides information regarding the distribution of “correct” awards for comparable claims, including the global mean, and, in turn, about the “correct” award for the subject claim.
he or she had brought an individual action,” putative class representatives will improve their ability to establish predominance under Rule 23(b)(3).68

In light of *Tyson Foods*, a court considering Rule 23(b)(3) certification is likely to focus more heavily and more explicitly on the reliability of representative evidence and statistical sampling generally. Although a statistical sample should be carefully scrutinized to detect, among other things, methodological deficiencies and issues regarding the cohesiveness of the class, sampling often improves the reliability of legal outcomes.

---

III. The Logic of Comparable-Case Guidance in the Determination of Awards for Pain and Suffering and Punitive Damages*

1. Introduction

Lawmakers and scholars have struggled to address the unpredictability of awards for pain and suffering and punitive damages. Jurors are provided with very little guidance in determining such awards, and courts lack objective standards to guide jurors and review their awards. Consequently, awards can vary wildly.

For example, in a well-known case involving a claim against BMW that the auto manufacturer had fraudulently sold cars as new after painting over parts that had suffered corrosion damage due to acid rain exposure while in transit to the United States, an Alabama jury awarded the purchaser compensatory damages of $4,000 and punitive damages in the amount of $4 million. However, in a materially identical case brought by another purchaser in the same court and before the same judge, but with a different jury, the jury awarded a similar level of compensatory damages, but no punitive damages at all. The jury found that BMW’s behavior did not rise to the level of reprehensibility deserving of punitive damages.2

---


1 The Alabama Supreme Court reduced the award to $2 million, and the United States Supreme Court ultimately reversed, holding that even $2 million in punitive damages was grossly excessive. See BMW of North America, Inc. v. Gore, 517 U.S. 559 (1996).

The awards in the BMW case, like others—such as a medical malpractice case resulting in a pain and suffering award of $100 million\(^3\) or a deceptive-cigarette-marketing case resulting in a punitive damages award of $28 billion\(^4\)—represents a stark deviation from Justice Holmes’s characterization of the law as a “systematized prediction.”\(^5\) As Judge Niemeyer commented regarding the $100 million award in *Evans*, “Because the jury was given no rule nor any rational criteria to apply in setting the amount of such an award, but told simply to do what it thought best, the jury responded with a perceived ‘measurement’ of the pain, which essentially amounted to an emotional response.”\(^6\) The jury award in *Evans*—an award ten times the amount requested by the plaintiff’s attorney, and one that the trial judge ultimately reduced to $3.5 million—suggests, at least, that such awards are unpredictable.\(^7\) Regarding the award in *Bullock*, Judge Niemeyer commented that “only emotion, not a rule of law, could justify imposing an award of $28 billion. To confirm this,

---


\(^5\) Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 458 (1897). Justice Holmes observed that “[p]eople want to know under what circumstances and how far they will run the risk of coming against what is so much stronger than themselves, and hence it becomes a business to find out when this danger is to be feared. The object of our study, then, is prediction, the prediction of the incidence of the public force through the instrumentality of the courts.” *Id.* at 457. *See also* Paul v. Niemeyer, *Awards for Pain and Suffering: The Irrational Centerpiece of our Tort System*, 90 VA. L. REV. 1401, 1402-04 (2004).

\(^6\) Niemeyer, *supra* note 5, at 1403.

\(^7\) *Id.* at 1403-04.
one need only consider the ‘mind’ of a legislative body developing a prospective rule of law to punish similar conduct.”

Indeed, numerous empirical studies have confirmed substantial anecdotal evidence that awards for pain and suffering and punitive damages are highly unpredictable.

Commentators have proposed numerous methods to address unpredictability, including various forms of award schedules that bind or guide a jury in its award determination in light of its findings regarding certain facts, such as the severity of a plaintiff’s injuries. Thus, “[s]chedules with categories based on injury severity typically provide the method of classification, and prior awards for injuries within each category provide a range of damages amounts.” That is, “[a]lthough reforms of this type differ in

8 Id. at 1409-10 (noting that the “California legislature has fixed the maximum fine for false advertising at $2500,” and that “Congress has fixed the maximum fine for corporations’ violations of federal offenses at $500,000 or twice the defendant’s gain or the victim’s loss”).

9 See 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) (concluding that there is “considerable variation in both juror and jury awards,” and that “[a] substantial portion of that variation is not predictable from measures of either background or attitudinal individual differences across jurors”); Randall R. Bovbjerg et al., Valuing Life and Limb in Tort: Scheduling “Pain and Suffering,” 83 NW. U. L. Rev. 908, 919-24 (1989) (concluding that “[a]lthough the median, and even mean, awards in a given category may be considered relatively reasonable, the seemingly uncontrolled variability of awards is cause for concern—similar to anxiety about drowning in a pool averaging only two feet in depth”); Oscar G. Chase, Helping Jurors Determine Pain and Suffering Awards, 23 Hofstra L. Rev. 763, 768-69 (1995) (citing studies); David W. Leebron, Final Moments: Damages For Pain and Suffering Prior to Death, 64 N.Y.U. L. Rev. 256, 259 (1989) (concluding that “tort awards for even this relatively simple area [of damages for 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”). See also Richard Abel, General Damages Are Incoherent, Incalculable, Incommensurable, and Inegalitarian (but Otherwise A Great Idea), 55 DePaul L. Rev. 253, 291-303 (2006). But see Yun-chien Chang et al., Pain and Suffering Damages in Personal Injury Cases: An Empirical Study, Univ. of Chicago Coase-Sandor Inst. for Law & Econ. Research Paper No. 749 (April 12, 2016), http://papers.ssrn.com/abstract=2741180 (concluding that “pain and suffering damages in Taiwan are to a large extent statistically and legally predictable”).

10 See infra Section 3.

their details, each proceeds from the premise that prior pain-and-suffering awards for similar cases provide the appropriate basis for computing the present award.”12 Ultimately, “[t]he jury or reviewing court determines where the plaintiff ’s injury falls on the schedule, and the schedule provides a range or specified amount that can be binding or nonbinding on juries or courts.”13

To the extent that these methods predetermine the award or range of awards, and to the extent that they bind the jury rather than guide it, they have been criticized as “eviscerat[ing] the various contributions that juries make to the civil justice system,” and as being “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.”14

A similar recommendation, proposed previously in various forms, involves “comparability analysis.” Using this approach, the court (perhaps by way of an adversarial process involving the litigation parties, and even the trier of fact) would first identify a universe of comparable cases. It would then provide the trier of fact with certain information regarding the awards in these cases in the context of a jury instruction or as expert testimony, and it would instruct the trier of fact to arrive at a damages determination in light of the evidence introduced in the case, and using the comparable-case information (or “prior-award information”) as guidance.15 Such methods are based, in

12 Id.
13 Id.
15 See infra Section 3. Note that there is substantial overlap between such methods and those involving scheduling—particularly to the extent that scheduling methodologies involve providing a fact-finder with scenarios and associated award values, based on comparable cases, as guidance in determining damages.
part, on empirical studies confirming that they are effective in controlling outlying awards and reducing award variability generally, even using prior awards to guide, rather than bind, the trier of fact.\footnote{See, e.g., Saks et al., supra note 15, at 249-55.}

For purposes of clarity—because the terms “damage schedule” and “comparability analysis” have been used in the literature to mean various things in various contexts—I will use the term “comparable-case guidance” (CCG) to refer to scheduling or comparability-analysis methods that fulfill three fundamental requirements: 1) information used as guidance must be derived from prior “comparable” cases (as opposed to, e.g., damage schedules predetermined arbitrarily by a legislative body); 2) comparable-case information must be considered by the trier of fact in particular (as opposed to, e.g., a reviewing court); and 3) comparable-case information must be used as guidance only (as opposed to, e.g., imposing a range or amount that is binding on the trier of fact).\footnote{Note that the analysis herein may apply to scheduling and comparability analysis more broadly; however, I focus on CCG methods in particular.}

While CCG methods allow for a case-by-case analysis that remains in the discretion of the trier of fact, the use of prior awards itself has been attacked based on three fundamental objections. First, “[i]f 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,” since “[b]y doing so, we may ensure that like cases are treated alike in that all involve

\textit{See also} Michael J. Saks et al., \textit{Reducing Variability in Civil Jury Awards}, 21 LAW AND HUMAN BEHAVIOR 243, 246 (1997); Bovbjerg et al., \textit{supra} note 9, at 953.
inappropriate damages awards.”18 Second, such methods “fail to address the fundamental issue of how one should initially assess the value of pain-and-suffering damages” or arrive at an appropriate punitive damages award.19 Third, the validity of such methods relies heavily on the presumption that a “correct” set of cases has been identified—that the prior cases identified are indeed materially “comparable” to the case at hand.20

These objections boil down to a fundamental concern: if our problem is the unpredictability of awards caused by a trier of fact’s inability to assess objectively the appropriate value of awards for pain and suffering or punitive damages, how is it beneficial to provide a trier of fact with information regarding damages awarded in prior cases that are separate and distinct from the present case and that presumably suffer from the same

18 Geistfeld, supra note 11, at 792 (commenting that “[t]his reliance upon past awards . . . represents the most problematic aspect of these reform proposals”). See Peter H. Schuck, Scheduled Damages and Insurance Contracts for Future Services: A Comment on Blumstein, Bovbjerg, and Sloan, 8 YALE J. ON REG. 213, 218 (1991) (“by 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”); see also Robert L. Rabin, The Pervasive Role of Uncertainty in Tort Law: Rights and Remedies, 60 DEPAUL L. REV. 431, 448-49 (2011) (“Scheduling proposals of this kind have been criticized by torts scholars such as Mark Geistfeld, who pointed out the seeming paradox in rejecting unstructured jury decision making in favor of a scheduled approach, which from a horizontal equity perspective takes arbitrary prior awards as the cornerstone for future awards, and from a vertical equity perspective takes the ordering of magnitude in past jury awards as an appropriate key for hierarchical sorting in the designated severity-categories for future awards”).

19 Ronen Avraham, Putting A Price on Pain-and-Suffering Damages: A Critique of the Current Approaches and A Preliminary Proposal for Change, 100 NW. U. L. REV. 87, 104 (2006) (discussing pain-and-suffering awards in particular, and citing W. Kip Viscusi, Pain and Suffering: Damages in Search of A Sounder Rationale, 1 MICH. L. & POL’Y REV. 141, 168 (1996) (“The key issue to be resolve[d] for pain and suffering schedules and scales is that establishing such arbitrary benchmarks does not resolve the more fundamental issue of how one should initially assess the value of pain and suffering damages”)).

20 See Logan, supra note 14, at 943-44 (“While [an approach allowing the fact-finder to consider a range of possible awards for guidance] would improve predictability, such an approach 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”).
arbitrariness that we wish to address in the present case? Seemingly, this would only compound the problem.

My aim in this article is to address these objections by explaining in simple but formal terms how, notwithstanding the foregoing objections, CCG methods reduce unpredictability and improve awards for pain and suffering and punitive damages generally by allowing for the sharing of relevant information across cases.

Section 2 discusses the importance of reducing the variability of awards for pain and suffering and punitive damages. Section 3 provides background regarding current methods and proposals for reducing award variability. Section 4 explains how CCG methods reduce unpredictability and improve awards for pain and suffering and punitive damages generally. Section 5 discusses a number of considerations for identifying a universe of prior cases and for distilling information from such cases for consideration by the trier of fact. Section 6 provides additional legal context for CCG methods and concludes.

2. The Importance of Reducing the Variability of Awards for Pain and Suffering and Punitive Damages

The Supreme Court has repeatedly emphasized the importance of maintaining fair and consistent awards for pain and suffering and punitive damages. In a recent case, for example, the Court stated:

The real problem, it seems, is the stark unpredictability of punitive awards. Courts of law are concerned with fairness as consistency, and evidence that the median ratio of punitive to compensatory awards falls within a reasonable zone, or that punitive awards are infrequent, fails to tell us whether the spread
between high and low individual awards is acceptable. The available data suggest it is not.\textsuperscript{21}

The Supreme Court continued by discussing the “inherent uncertainty of the trial process” and the resulting inconsistency among awards in cases with similar facts.\textsuperscript{22} In examining unpredictability as a matter of policy, rather than of constitutional significance,\textsuperscript{23} the Court emphasized that the unpredictability of high punitive damage awards is “in tension with the function of the awards as punitive.”\textsuperscript{24} It commented:

Thus, a penalty should be reasonably predictable in its 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. See The Path of the Law, 10 Harv. L.Rev. 457, 459 (1897). 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. Cf. Koon v. United States, 518 U.S. 81, 113 (1996) (noting the need “to reduce unjustified disparities” in criminal sentencing “and so reach toward the evenhandedness and neutrality that are the distinguishing marks of any principled system of justice”). The common sense of justice would surely bar penalties that reasonable people would think excessive for the harm caused in the circumstances.\textsuperscript{25}


\textsuperscript{22} Id., 554 U.S. at 500-01 (quoting BMW of North America, Inc. v. Gore, 646 So.2d 619, 626 (1994)).

\textsuperscript{23} The Supreme Court highlighted that “the Court’s response to outlier punitive-damages awards has thus far been confined by claims at the constitutional level, and our cases have announced due process standards that every award must pass.” Id. at 501. In Baker, however, the Court examined a jury award for punitive damages “for conformity with maritime law, rather than the outer limit allowed by due process.” Id. at 501-02. In acting “in the position of a common law court of last review,” id. at 507, it considered punitive damages not with respect to “their intersection with the Constitution,” but rather in relation to the “desirability of regulating them as a common law remedy.” Id. at 502.

\textsuperscript{24} Id. at 502 (emphasis added).

\textsuperscript{25} Id. at 502-03.
The Supreme Court thus emphasized the point that predictability, consistency, and fairness are fundamental to the deterrence objectives underlying punitive damages. This sentiment and others associated with the harms of unpredictability have been echoed repeatedly by lower courts and scholars.

For example, in the case of *Payne v. Jones*, the Second Circuit explained the purposes underlying the device of remittitur by emphasizing the harmful effects of variability and outlying awards, including those associated with over-deterrence:

Apart from impairing the fairness, predictability and proportionality of the legal system, judgments awarding unreasonable amounts as damages impose harmful, burdensome costs on society. As an initial matter, an excessive verdict that is allowed to stand establishes a precedent for excessive awards in later cases. The publicity that accompanies huge punitive damages awards, *see, e.g.*, Henry Weinstein, *Philip Morris Ordered to Pay $28 Billion to Smoker*, L.A. Times, Oct. 5, 2002, will encourage future jurors to impose similarly large amounts. Unchecked awards levied against significant industries can cause serious harm to the national economy. Productive companies can be forced into bankruptcy or out of business. Municipalities can be drained of essential public resources. The threat of excessive damages, furthermore, drives up the cost of insurance premiums, deters both individuals and enterprises from undertaking socially desirable activities and risks, and encourages overspending on “socially excessive precautions” that “cost[ ] more than the reduction of harm produced by [them].” A. Mitchell Polinsky & Steven Shavell, *Punitive Damages: An Economic Analysis*, 111 Harv. L. Rev. 869, 879 (1998). The prices of goods and services will rise, and innovation will be inhibited. See id. at 873.26

Courts have similarly emphasized the need to “minimize the arbitrary variance in awards bound to result from [the] throw-up-the-hands approach” that courts regularly use in determining awards for pain and suffering.27 Courts and scholars have recognized the

---

26 *Payne v. Jones*, 711 F.3d 85, 94 (2d Cir. 2013).

27 *Jutzi-Johnson v. United States*, 263 F.3d 753, 759 (7th Cir. 2001).
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.”

In Geressy v. Digital Equip. Corp., Judge Weinstein explained the court’s decision to consider prior-award information. He discussed concern by the courts and legislature for the “virtually unbridled discretion” of juries in awarding damages for pain and suffering, for which there is “currently no meaningful way to measure such non-quantifiable losses monetarily.” As Professor Oscar Chase explained:

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; that is, it makes it harder to settle cases, thus adding unnecessary transaction costs to the tort system, and delaying payment to needful plaintiffs. Unpredictability also leads to inefficiencies because of over- or under-precautions by affected industries and insurers.

Thus, whether for purposes of fairness, deterrence, or another objective, courts recognize the importance of generating consistent and predictable damage awards. Indeed, reducing variability is fundamental to achieving reliable legal outcomes. In Section 4, I consider variability in the context of error generally, and discuss in more detail what is meant by “reliable legal outcomes.” First, however, I describe current methods and proposals for addressing award variability.

28 Id.


30 Chase, supra note 9, at 769.
3. **Current Methods and Proposals for Addressing Award Variability**

**Additur and Remittitur**

Courts use the procedural devices of additur and remittitur to increase or decrease an award found to be insufficient or excessive. For example, a defendant may argue for a new trial based on the excessiveness of the jury's award. If the judge agrees, he may offer the plaintiff to reduce the award (remittitur) by some amount rather than proceeding with a new trial.³¹ Appellate courts may also modify awards and address challenges to additur and remittitur.³²

Although the devices of additur and remittitur can, *in theory*, address the high levels of variability associated with awards for pain and suffering and punitive damages, in practice they cannot. First, additur and remittitur are primarily used to address excessive awards, and are infrequently used to adjust inadequate awards.³³ Second, these procedures are generally reserved for only the most extreme cases—they are used to address extreme outliers that are deemed incorrect, rather than address variability in general.³⁴ Third, courts lack consistent and principled procedures for arriving at better award determinations than juries. As Judge Posner opined in *Jutzi-Johnson v. United States*, “[most

---


³² *Id.* Both state and federal courts use the additur and remittitur procedures; but the Supreme Court has distinguished additur from remittitur and held that the former procedure is violative of the defendant’s Seventh Amendment right to a trial by jury. *See Dimick v. Schiedt*, 293 U.S. 474 (1934).

³³ See Baldus et al., *supra* note 31, at 1119-20.

³⁴ *See id.* at 1118-20.
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." Further, courts
employ these devices infrequently; and widespread replacement of jury awards with
judicial determinations would be problematic with respect to the Seventh Amendment and
fundamental principles of tort law.

**Damage Caps.** Similarly, damage caps, which place upper limits on damage awards
or certain types of damage awards, are widely recognized as a particularly poor method for
addressing variability. They give rise to a range of problems inherent in capping damages

---

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

36 See Baldus et al., supra note 31, at 1119-21 (citing Michael G. Shanley & Mark A. Peterson, *Posttrial Adjustments to Jury Awards* vii-viii, 43-47, RAND: THE INST. FOR CIVIL JUSTICE (1987) http://www.rand.org/content/dam/rand/pubs/reports/2007/R3511.pdf, and others). Baldus et al. note that “[o]n the basis of post-trial information in the 880 1982-84 California and Cook County, Illinois cases reported in [the Shanley and Peterson study], we estimate that in cases involving a damages award, there was a remittitur or new trial 6% of the time and an additur or a negotiated increase in the award in 2-3% of the cases." See also 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, 503 (2006).

37 Damage caps have served as the primary tool of tort reform aiming to address variability. Legislation for caps boomed during heightened calls for tort reform in the 1970s and 1980s, and by 1987, twenty-three states had instituted caps ranging from $150,000 to $1,000,000. Sanders, supra note 36, at 510. See also Stephen D. Sugarman, *A Comparative Law Look at Pain and Suffering Awards*, 55 DePaul L. Rev. 399, 399-400 (2006). Since then, numerous other states have adopted caps in various contexts as well. Sanders, supra note 36, at 510. Statutory caps may be applied to specific types of damages, such as punitive damages or general damages, or damage awards generally. See Baldus et al., supra note 31, at 1121-23. They may be applied in particular contexts, such as medical malpractice. See generally Kathryn Zeiler, *Turning from Damage Caps to Information Disclosure: An Alternative to Tort Reform*, 5 Yale J. Health Pol’y, L & Ethics 385 (2005). And they appear in a variety of forms. For example, juries can be informed of damage caps explicitly, or a cap can be applied by the court only after a jury has exceeded it. In general, juries are not informed of caps; but, regardless, jurors may be aware of such measures. See Saks et al., supra note 15, at 245. Under some proposals, a judge may be permitted to reduce awards below the quantity provided by the legislation; in others, a judge may be permitted only to reduce an award to the established quantity. See Colleen P. Murphy, *Statutory Caps and Judicial Review of Damages*, 39 Akron L. Rev. 1001, 1002-08 (2006).
(or certain types of damages) outright, independent of the circumstances of the case or harm suffered by the plaintiff.

First, caps address only the most extreme cases, and only excessive awards. Second, caps are likely to play a role primarily in cases in which the harm is very severe. If, for example, there is a cap of $1 million, the cap may not affect an individual who suffered a broken nose, but would greatly affect an individual who suffered paralysis. The cap is unlikely to address variability in the former case, while possibly causing an inappropriately low award in the latter. Thus, aside from biasing the award, caps distort incentives to litigate tort claims based on severe harm. Further, caps can cause suboptimal risk-taking by diluting the deterrence effect of punitive damages and tort law generally. Third, a number of studies have concluded that caps may in fact exacerbate the variability problem rather than address it. In particular, jurors who are aware of a cap may anchor to it, and their awards may gravitate to it. “Thus, those with the most severe injuries and losses will be unfairly deprived of compensation by caps. In addition, by serving as a psychological anchor, caps appear likely to further exaggerate the error of overcompensating those with smaller losses.”

38 See Baldus et al., supra note 31, at 1121-23.

39 See Sanders, supra note 36, at 509-11. Constitutionality issues may arise as well. Opponents of caps have argued, for example, that caps are violative of a plaintiff’s right to trial by jury. See Zeiler, supra note 37, at 387.

But, notwithstanding the foregoing problems, damage caps are used, due, in part, to their administrative convenience and their ability to control insurance rates.\textsuperscript{41}

**Comparable-Case Information.** Finally, a number of courts and commentators have proposed methods involving the use of information regarding awards in prior “comparable” cases. Such proposals have appeared in various forms. One set of approaches builds on courts’ regular use of comparability review, whereby trial and appellate courts consider awards in comparable cases in their review of a trier of fact’s award for excessiveness.\textsuperscript{42} These proposals attempt to develop a more structured framework for the court’s comparative analysis—for example, instituting a method by which a court identifies a universe of comparable claims; examines whether the current award “deviates materially from what would be reasonable compensation” (to use language from a New York statute upon which such proposals seem to build);\textsuperscript{43} and, if so, determines a suitable adjustment of the jury award.\textsuperscript{44}

These review-based methods have the advantage of involving only modest modifications to current well-accepted methods, but they are not ideal. Specifically, they incorporate comparable-case information on review rather than in the award directly. As such, they only focus on outlying awards rather than on variability in general. Moreover,

\textsuperscript{41} See Sanders, supra note 36, at 511.

\textsuperscript{42} See Baldus et al., supra note 31, at 1134.

\textsuperscript{43} N.Y. C.P.L.R. 5501(c) (McKinney 1995).

\textsuperscript{44} See Baldus et al., supra note 31, at 1134. See also Sanders, supra note 36, at 503.
regular use of such methods arguably leads to constitutionality and tort issues associated with the replacement of a jury’s discretion with that of the judge.⁴⁵

As indicated in the Introduction, another set of proposals uses various types of schedules to bind or guide the trier of fact in its determination of certain types of damage awards.⁴⁶ These proposals, to the extent they involve binding a trier of fact to predetermined values or ranges of values, have been criticized as impinging on the trier of fact’s discretion, and as “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.”⁴⁷ To the extent that such methods bind rather than guide the trier of fact, and to the extent that they involve award values that are predetermined, for example, by a legislative body in advance of the case, such methods are arguably problematic.

Finally, a number of proposals—overlapping with the category above regarding schedules—involve comparability analysis, whereby the court (perhaps by way of an adversarial process involving the litigation parties, and even the trier of fact) identifies a set of materially “comparable” cases, provides the trier of fact with certain information regarding the awards in these cases in the context of a jury instruction or presented as

⁴⁵ See Colleen P. Murphy, Judicial Assessment of Legal Remedies, 94 NW. U. L. REV. 153, 192 (1999); J. Patrick Elsevier, Out-of-Line: Federal Courts Using Comparability to Review Damage Awards, 33 GA. L. REV. 243, 258-59 (1998) (“Comparability analysis [on review] requires a court to engage in several subjective, factintensive inquiries . . . Yet, by asserting that comparability provides an objective framework to review compensatory damage awards, courts are able to substitute their own assessment of what the appropriate damage award should be, while skirting the Seventh Amendment’s proscription against reexamination of issues of fact”).

⁴⁶ See, e.g., Bovbjerg et al., supra note 9, at 945.

⁴⁷ Logan, supra note 14, at 943.
expert testimony, and instructs the trier of fact to determine an appropriate damages award in light of the evidence introduced in the case, and using the comparable-case information as guidance.\textsuperscript{48} As suggested in the Introduction, such methods—and CCG methods in particular—avoid many of the problems caused by other methods, such as damage caps and binding or predetermined schedules. These methods allow for a flexible case-by-case analysis in the discretion of the trier of fact, while still providing substantial guidance. Nevertheless, such methods have been attacked on the grounds that, instead of addressing the fundamental issue of assessing the value of pain-and-suffering damages or an appropriate punitive damages award, they may \textit{compound} the unpredictability of damage awards by providing the trier of fact with information regarding damages awarded in separate and distinct prior cases that presumably suffer from the same arbitrariness that the court aims to address in the present case.\textsuperscript{49} In the following sections, I address these objections.

4. \textbf{The Logic of CCG}

In this section, I explain why, notwithstanding the objections delineated above, CCG methods reduce unpredictability and improve the reliability of awards for pain and suffering and punitive damages generally. I begin by discussing what is meant by

\textsuperscript{48} See, \textit{e.g.}, Chase, \textit{supra} note 9, at 777-90 (proposing that courts inform jurors of “the range of awards made by other juries in the same state for such damages during a contemporaneous time period,” to be provided as non-binding guidance in the court’s charge to the jury in the form of a “chart constructed to allow comparison with roughly similar cases in which plaintiffs’ verdicts were recovered.” \textit{Id.} at 777.). \textit{See generally} Baldus et al., \textit{supra} note 31, at 1123-24; Saks et al., \textit{supra} note 15, at 246; Sanders, \textit{supra} note 36, at 506-7; \textit{Jacob A. Stein, Stein on Personal Injury Damages Treatise} \textsection{8:32} (3d ed. 2015) (discussing American Bar Assoc., \textit{Report of the Action Commission to Improve the Tort Liability System} 10-15 (1987)); Jutzi-Johnson v. United States, 263 F.3d 753, 758-61 (7th Cir. 2001).

\textsuperscript{49} See Geistfeld, \textit{supra} note 11, at 792; Avraham, \textit{supra} note 19, at 104; Logan, \textit{supra} note 14, at 943-44.
“reliability.” I then introduce the role of CCG in reducing variability and error in, and thus improving the reliability of, legal outcomes.

4.1 Bias and Variance: The Elements of Error

Section 2 makes clear that controlling the variability of awards for pain and suffering and punitive damages is important. But reducing variability does not necessarily mean improving the award. For example, it is likely unwise to encourage jurors to anchor to an arbitrary value, notwithstanding associated variability benefits. In fact, variability could be zero if the court were to dictate an award without regard to the particulars of the case.

Courts and scholars are correct to be concerned about the high variability associated with awards for pain and suffering and punitive damages. But good policy requires steps toward reducing such variability only insofar as they improve awards generally.

Thus, for purposes of considering the concepts of “error” and “accuracy” in the context of damage awards, assume that there is some “correct” award that could, in theory, be determined as a function of perfect information regarding the state of the world, including the facts of the case and applicable law and norms. Of course, we neither have perfect information nor know the correct award. Instead, the court asks a jury to arrive at an award, which will serve as an estimate of the correct outcome.

---

50 See Hillel J. Bavli, Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation, 14 LAW, PROBABILITY & RISK 67, 74-78 (2015). We can similarly consider a range or distribution of “correct” awards that reflects, for example, uncertainty regarding the law. Id. at 74 n.24.
More concretely, let us consider the correct award in a given case—say, for example, the BMW case described in the Introduction—to be 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.).\footnote{Id. (citing 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)). For simplicity, I ignore the potential for bias. As noted in Bavli, supra note 47, “[i]t may be more intuitive to consider the concept of a ‘correct’ verdict in the context of a criminal trial. Consider, for example, the O.J. Simpson murder trial. Polls show that over 50% of Americans believe that the jury arrived at an incorrect verdict. Implicit in the public’s disagreement with the verdict is an assumption that there exists a ‘correct’ outcome. The framework described above is intuitive: had the jury known the true facts of the case, and had there been no ambiguity regarding the application of the law to the facts of the case, the jury would have arrived at the correct conclusion. But given ambiguity regarding either the facts of a case or the state of the law, it is unclear whether a criminal defendant in fact satisfied the elements of the crime charged; and the jury must arrive at a verdict—‘guilty’ or ‘not guilty’—which serves as an estimate of the correct outcome, ‘guilty’ or ‘not guilty.’” Id. at 74 n.29. Similarly, the determination of a civil damages award (or, e.g., a prison sentence) can be understood as an estimate of a correct outcome. Arguably courts implicitly acknowledge this characterization when, for example, a court finds a jury award to be excessive or inadequate. Finally, the formulation of the correct award as the mean of the population of possible awards over repeated adjudications is intended to capture, e.g., the state of the law as understood by various judges across the relevant society, and the norms of the time and facts of the case as understood by various combinations of jurors in the relevant society.} A single trial thus generates a sample from the population and an estimate of the correct award.\footnote{See id. at 74-75; Saks & Blanck, supra note 51, at 833-34.} Call the actual award an “estimate,” and the procedure that generates the estimate an “estimator.”\footnote{Id. The estimator can be understood as the legal procedure—including the rules governing the trial, the role of the judge, the jury deliberation, etc.—that generates the legal outcome, which, in our example, is the jury’s punitive damages award.} We can then define “error” in terms of distance, and (equivalently) “accuracy” in terms of proximity, between the estimate and the correct award.\footnote{See id. at 74-78.}

I define the reliability of a legal procedure as the accuracy of the outcome (here, the award) that we can expect by following the procedure. If we expect that a certain legal
procedure will produce an accurate award, we say that the procedure (or, for simplicity, the award) is “reliable.”

In statistics there are two sources of error: bias and variance. If an estimator 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 estimator generates on average and the correct value. Bias is therefore a source of error. Note that it is possible that an unbiased estimator will never generate the correct value—to be unbiased is only to generate the correct value in expectation, or on average. Additionally, although unbiasedness is generally understood as a good characteristic for an estimator to have, it does not indicate lack of error, since the values generated by the estimator can vary wildly around the correct value. For example, if the correct punitive damages value in the BMW case above 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, $100,000, but 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.

Thus, the second source of error is “variance,” which is a measure of dispersion. In particular, if an estimator entails a high level of variance, then it will generate estimates that are highly dispersed around its mean, or average, value. If the estimator entails a high

---

55 See generally id.
level of variance, but is unbiased, then it will generate estimates that are highly dispersed around the correct value. In this case, we say that the estimator is “unbiased,” but that it lacks “precision.” If the estimator is “precise” but “biased,” then it generates values that are tightly centered around the wrong value—an undesirable circumstance. If an estimator is “precise” and “unbiased,” then it will generate estimates that are close in proximity to the correct value, and we say that it is “accurate.”

Thus, in the BMW example above, the awards $90,000, $95,000, $105,000, and $110,000 reflect greater precision than $0, $50,000, $150,000, and $200,000. And, the awards $100,000, $100,000, $100,000, and $100,000 reflect even greater precision.

More formally, let $\alpha$ be the “correct” award and $\hat{\alpha}$ a random variable defined by the actual award. Let $E(\hat{\alpha})$ be the expectation of $\hat{\alpha}$. In other words, $\hat{\alpha}$ will equal $E(\hat{\alpha})$ on average; and if we repeat the trial many times then the average of the punitive damage awards will be approximately $E(\hat{\alpha})$.

Bias is defined as the difference between the expectation of the estimator $\hat{\alpha}$ and the correct award $\alpha$ (the object we are trying to estimate). More formally $Bias = E(\hat{\alpha}) - \alpha$, and it is said that the estimator $\hat{\alpha}$ is “unbiased” if $E(\hat{\alpha}) = \alpha$.

Additionally, let $V(\hat{\alpha})$ be the variance of $\hat{\alpha}$, where variance is defined in the standard way. That is,

$$V(\hat{\alpha}) = E((\hat{\alpha} - E(\hat{\alpha}))^2) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\alpha}_i - E(\hat{\alpha}))^2.$$
The variance of the realized awards is thus the average of the square differences between the awards and the average of the awards. For example, the variance of $0, $50,000, $150,000, and $200,000 is $rac{1}{4} ((0 - 100,000)^2 + (50,000 - 100,000)^2 + (150,000 - 100,000)^2 + (200,000 - 100,000)^2)$. The standard deviation of $\hat{\alpha}$ is the square root of the variance: $SD(\hat{\alpha}) = \sqrt{\text{Var}(\hat{\alpha})}$.

Now let us formally define error in terms of bias and variance. There are many ways to measure error (just as there are many ways to measure, for example, dispersion). For instance, we might define error by the raw differences between the estimates and the correct value. But just as it is inconvenient for scientists to define dispersion by calculating the raw differences between the estimates and the mean of the estimates, similar inconveniences arise from defining error in terms of raw differences. It turns out that a convenient and often suitable measure of error is the well-accepted mean squared error ($\text{MSE}$), defined as $\text{MSE}(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2]$. In words, we find the expectation of the square difference between the estimator and the correct value.

Now, using $\text{MSE}$ as our measure of error, it is easy to show that:

$$\text{MSE}(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2] = \text{Variance} \times \text{Bias}^2$$

That is, error (defined here as MSE) can be separated into two elements: bias and variance.\(^57\)

---

\(^56\) For example, using such differences results in the “canceling” of positive and negative values.

\(^57\) See Sharon L. Lohr, SAMPLING: DESIGN AND ANALYSIS, §2.2 (2d ed. 2010), for a mathematical proof, and a lengthier discussion, of the bias-variance decomposition.
4.2 Reducing Variability and Error with Prior-Award Information

Let us view a civil case as an estimation problem within the framework described above. That is, for every claim let us assume that there is a “correct” damages award (or, in particular, a “correct” award for pain and suffering or punitive damages) $\alpha$, defined as the mean of the population of awards that would result from repeated adjudications, as described above (and that would reflect, for example, perfect information regarding the applicable law and the facts of the case).\(^5\) As discussed, we cannot know the correct award and must therefore arrive at a suitable estimate $\hat{\alpha}$ of the unknown correct award.

Thus, let us assume (for simplicity rather than necessity) that, on average, the estimate will equal the correct award, but that it entails some degree of variability $\sigma^2$ around the correct award. That is, $\hat{\alpha}$ is distributed with mean $\alpha$ and variance $\sigma^2$.

Notationally, $\hat{\alpha} \sim (\alpha, \sigma^2)$

Our goal is to arrive at a damages estimate that is close to the correct award—that is “accurate” in the sense of minimizing error, as defined above by $MSE(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2]$.

In statistics, and the sciences generally, determining a good estimator for an unknown quantity is very important and is the subject of much research. Let us use an analogy to better understand the problem and possible solutions. Consider a study by statisticians Bradley Efron and Carl Morris, in which they recorded batting average data for each of eighteen major league baseball players through his first forty-five official at bats of

---

\(^5\) See Bavli, supra note 50, at 74-78. We can define the “correct” award using other measures of central tendency, such as the median, depending on how we want to characterize it. The mean, for example, is more sensitive than the median to extreme awards. For simplicity, our analysis in this article focuses only on the mean.
the 1970 season. Suppose our goal is (as their goal was) to estimate the batting averages that would emerge for each player during the remainder of the season, using only the data collected. Let us focus on one player, Roberto Clemente, whose batting average for his first 45 at bats was .400.

Initially, it might seem elementary that a player’s batting average for the first forty-five at bats would serve as the best estimator for his batting average for the remainder of the season. Thus, our estimate of Clemente’s remainder-of-the-season (or “after-forty-five”) batting average would be .400. But it can be shown mathematically and empirically that this is not in fact the best estimator. In particular, this estimator—Clemente’s batting average for his first forty-five at bats—can be improved by incorporating information regarding the batting averages for the first forty-five at bats of the other seventeen players. But why should this be? Why should the batting averages of Munson or Kessinger, who had “first-forty-five” batting averages of .156 and .289, respectively, have any influence on our estimate of Clemente’s after-forty-five batting average? Using other players’ first-forty-five batting averages would seem obvious in the absence of Clemente’s first-forty-five average; but we have Clemente’s first-forty-five batting average—and what could be better than that?

The explanation lies in the information we obtain from the first-forty-five batting averages of the other seventeen players. First, suppose we know nothing about batting

---


60 Id. at 313.

61 See id. at 312-14.
averages. It seems intuitive that, in estimating Clemente’s after-forty-five batting average, it is helpful to know whether his first-forty-five batting average, .400, is “low,” “high,” or “average.” In particular, it seems helpful to know that .400 is an extremely high batting average. As an extreme example, if we are told that, in a given period, a particular batting average is the highest ever achieved by that player, or in baseball in general, would we still use this average as our best estimate of the player’s average in the next period? We would be wise to incorporate this information, since the player is unlikely to break baseball’s all-time record twice in a row. More specifically, the usefulness of other-player information can be explained, in part, by considering the old idea (due to Sir Francis Galton) of “regression toward the mean,” which, in its simplest form, states that if we obtain a relatively extreme measurement, then upon re-measurement, we are likely to obtain a new measurement closer to average.62 The reason is simple: an extreme measurement can be attributed to a combination of two things—the characteristics of the thing measured (e.g., the skill of a baseball player who was in fact a far superior batter than his peers) and random variation, or “luck.” To the extent that the extreme measurement is due to randomness, it is likely to be less extreme upon re-measurement. Similarly, information regarding other players’ performance provides context for Clemente’s performance, and an estimator that incorporates such information is likely better than one based solely on the player’s individual performance.63


63 We can consider two sources of variability: there is variation among the players (some are better than others, for example) and, given a particular player (and given his mean batting average in particular), there is variation in the player’s batting average. “Regression toward the mean” thus operates in two ways: First,
In short, if we believe that the players’ batting averages are somehow bound together, we can make significant estimation improvements by incorporating other-player information.\textsuperscript{64}

Returning to the legal context, the value of a damages award (or, specifically, an award for pain and suffering or punitive damages) can be understood as attributable to 1) the application of the law to the particular facts of the case, and 2) random variation. In order to gain a deeper understanding of the benefits of CCG, let us model an award “hierarchically” to account for both of the foregoing elements. As above, we assume that the actual award $\alpha$ is distributed with mean $\alpha$ and variance $\sigma^2$. But now, let us assume that $\alpha$ is itself random rather than constant, and that it is distributed with mean $\mu$ and variance $\tau^2$. Notationally, $\alpha \sim (\alpha, \sigma^2)$ and $\alpha \sim (\mu, \tau^2)$ In words, the hierarchical model incorporates variability on two levels. First, the upper level ($\alpha \sim (\mu, \tau^2)$) indicates that cases are heterogeneous—that there can be factual variability across cases, and that the present case may be materially different from prior cases. I refer to this form of variability, $\tau^2$, as “claim variability.”\textsuperscript{65} Second, the lower level ($\alpha \sim (\alpha, \sigma^2)$ indicates that, given the facts of any

\textsuperscript{64} Efron & Morris discuss conditions under which estimation improvements can be made even when the estimands—the things we wish to estimate—are not bound together. See generally Efron & Morris, supra note 59; Bradley Efron & Carl Morris, \textit{Stein’s Paradox in Statistics}, 236 \textit{Scientific American} 119 (1977).

\textsuperscript{65} See Bavli, supra note 50, at 74-78.
single case (for example, the present case), there is variability (i.e., randomness) in the determination of damages. I refer to this form of variability, $\sigma^2$, as “judgment variability.”

Again, our goal is to arrive at a good estimate $\hat{\alpha}$ of $\alpha$. On one hand, we want the estimate to reflect the first form of variability—that meaningful factual differences (e.g., between the present case and prior cases) should result in different awards. On the other hand, we want to minimize the second form of variability—the randomness associated with the determination of damages in a given case. So how does incorporating prior-award information help? Unfortunately, since we generally do not know the correct damages award, it is impossible to distinguish the former form of variability from the latter. In other words, a tradeoff arises between minimizing bias—that our estimated damages award should, on average, be as close to the correct damages award as possible, reflecting the former form of variability—and minimizing variance, the degree of randomness associated with the estimated award. Thus, in a sense, incorporating prior-award information as guidance in determining a damages award allows the trier of fact (implicitly) to strike a balance between bias and variance so as to minimize error. By accepting the possibility of introducing some bias due to material differences between the present case and prior cases, it is possible to gain far more, in terms of reliability, due to the reduction in award variability that follows from such guidance.

Furthermore, the incorporation of prior-award information should not be viewed as suffering from circularity—that is, as a method that is defeated by its reliance on

---

information that entails the same flaws that the method is itself attempting to address. Rather, such methods should be understood as facilitating the improvement of awards by allowing for information sharing across cases. In the framework described, for example, an ideal method for reducing error from variability would generate a damages award by repeatedly adjudicating a case many times independently and using the mean of the repeated adjudications as the ultimate award. But such methods are generally impractical. Methods involving the incorporation of prior-award information can be understood as efficient alternatives. Such methods aim to control variability through information-sharing across different cases, rather than across replications of the same case. The cost of efficiently reducing judgment variability, however, is the potential for bias arising from claim variability.

In order to gain a better understanding of the effects of prior-award information on error, let us consider two extreme scenarios. First, suppose that the prior-award information is “dogmatic” in the sense that it entirely dominates the damages award without leaving room for any influence from the facts of the present case itself—for example, suppose the jury decides dogmatically simply to apply the average of the previous awards as its award determination, without consideration of the facts of the current case. In this case, there would be no error from variability, since the average of the previous damage awards is constant; but to the extent that the average of the prior awards is different from the correct current award \( \alpha \), there would be substantial error from bias. On the other hand, suppose that the prior-award information has no influence on the current damages award—for example, suppose the jury decides dogmatically that it will arrive at
an award determination based on the facts of the case alone, without influence from prior
damage awards. In this case, assuming jury damage determinations are initially \( i.e., \)
without influence from prior awards) unbiased—that is, they achieve the correct award \( \alpha \)
on average—then there would be no error from bias, but substantial error from variability.

CCG improves the reliability 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.
The quality of the balance is determined by the strength of the method for identifying
comparable cases and distilling information for consideration by the trier of fact.

To be sure, let us consider two important issues. First, assuming an “appropriate”
set of prior cases, what is the ideal level of influence that should be afforded to prior-award
information? Second, how can we be certain that the prior cases identified are sufficiently
comparable to the present case so as to improve reliability?

The first issue—the ideal influence of prior-award information, assuming an
appropriate set of prior cases—depends on the judgment variability \( (\sigma^2) \) of the award in
the present case and the claim variability \( (\tau^2) \) of the prior cases. Higher judgment
variability suggests weaker information obtained from the present case and greater
influence of the prior awards; higher claim variability suggests weaker information
obtained from the prior cases and greater influence of the present case.

For example, consider a tort case in which the evidence against the defendant is
overwhelming and the damages incurred by the plaintiff are clear. Assume that the
evidence is sufficiently strong that, if the case were litigated ten times independently, the
damage awards in all ten replications would be approximately the same—say, $100,000. In this case, the evidence is so great and the judgment variability so low (almost zero) that there is little to be gained from introducing prior-award information and the possibility of bias. On the other hand, consider a case in which, although there may be strong evidence regarding the facts of the case, there is only weak evidence regarding the appropriate damages award—as is almost always the case for awards for pain and suffering and punitive damages. If the case were repeatedly adjudicated ten times independently, the damage awards across the ten replications would vary wildly. In this case, the damages award is highly variable and there is much to be gained by introducing prior-award information. Further, the prior-award information is especially beneficial if claim variability is low and prior awards provide clear information. For example, assume that the facts and legal issues in the present case are clear, and that the court confidently identifies a set of ten comparable cases that involve similar facts—for example, similar injuries arising from similar circumstances. Assume further that the damages awarded in the ten cases are very similar—say, approximately $50,000. In this case, in which judgment variability is high and claim variability is low, prior-award information should have a high degree of influence.67

67 More formally, it can be shown that, if, as above, the damages award \( \hat{\alpha} \) is distributed with mean \( \alpha \) and variance \( \sigma^2 \), and the “correct” damages award \( \alpha \) is distributed with mean \( \mu \) and variance \( \tau^2 \) (thus, if we have judgment variability \( \sigma^2 \) and claim variability \( \tau^2 \)), then the optimal damages award weights current case information by \( 1/\sigma^2 \) and prior-award information by \( 1/\tau^2 \). In other words, the ideal influence of the prior awards (and particularly the mean of the prior awards) is inversely proportional to the claim variability \( (\tau^2) \) of the prior awards relative to the judgment variability \( (\sigma^2) \) of the award in the present case. See generally Efron & Morris, supra note 59; W. James & Charles Stein, Estimation with Quadratic Loss, 1 Proceedings of the Fourth Berkeley Symposium 361 (1961). Note that since we cannot, in practice, know the value of \( \mu \), the mean of the correct awards, we instead estimate \( \mu \) using the mean of the prior awards, and
It is worth reemphasizing here that it is incorrect to conclude, as numerous commentators have, that, because prior awards may suffer from the same arbitrariness that we aim to address in the present case, CCG would be ineffective in reducing (or, worse, would compound) award variability. Even assuming the (not-unlikely) scenario that judgment variability is approximately equal across cases (as I have assumed above), the foregoing analysis makes clear that the ideal influence of prior-award information can nevertheless be very high. In other words, CCG may nevertheless significantly reduce variability and improve reliability. This makes sense, since CCG allows for the sharing of information across numerous (although variable) cases; it provides significant guidance where relevant information is otherwise scarce.

Consider an extreme example in which judgment variability is extremely high for all cases (the prior cases as well as the present case) and claim variability is extremely low. Here, although prior awards suffer from the same extreme arbitrariness that we aim to address in the present award, a court would benefit immensely from CCG methods. Specifically, because claim variability is extremely low, the prior-award information obtained here would be approximately equivalent to prior-award information that could be obtained from the “ideal” procedure described above, whereby a damages award is calculated by averaging over awards obtained from repeated independent adjudications (with different juries, judges, lawyers, presentations of the evidence, etc.) of a single case.

\[
\alpha = \frac{\frac{\sigma^2}{\bar{x}^2} + \frac{\sigma^2}{\bar{z}^2}}{\frac{\sigma^2}{\bar{x}^2} + \frac{\sigma^2}{\bar{z}^2}}
\]

“shrink” toward this value. We can estimate the variance terms as well. Notationally, the “shrinkage estimator” can be written as \( \alpha = \frac{1}{\frac{\sigma^2}{\bar{x}^2} + \frac{\sigma^2}{\bar{z}^2}} \).
the case at hand. The awards obtained from repeated adjudications would vary considerably, since judgment variability is extremely high. But the ultimate award, calculated by averaging all of the awards obtained from the repeated adjudications, is extremely precise—it is subject to almost no variability.

As suggested above, methods involving repeated adjudication are ideal but costly. CCG methods are efficient alternatives that come at the cost of possible bias that may arise due to “misalignment”—where the mean of the correct awards in the prior cases is different from the correct award in the subject case—and perhaps due to claim variability, or heterogeneity in the set of prior cases. In our current example, however, bias is not a problem, because there is no misalignment problem and claim variability is assumed to be extremely low. Thus, although the prior awards are subject to extremely high judgment variability, in the aggregate they allow for an extremely reliable outcome through the sharing of information across cases. Furthermore, we can relax the assumption of extremely low claim variability. Higher claim variability may cause bias; but the harm to reliability associated with any introduction of bias is often well outweighed by the reliability benefits associated with the reduction in judgment variability. In any case, the introduction of claim variability does not affect the conclusion that CCG methods are effective in reducing judgment variability and improving reliability notwithstanding the potential arbitrariness associated with the prior awards.

---

68 See infra note 81.
The second issue—the reliability benefits of CCG in light of the *actual* comparability of the prior awards—depends on whether the potential for bias is outweighed by the benefits of reducing variability. This is a valid concern in the sense that the reliability benefits of CCG are indeed dependent on the material comparability of the prior cases to the present case.

For two reasons, however, the likelihood of identifying a set of cases sufficiently inappropriate, relative to the present case, to *harm* reliability is very low. First, because awards for pain and suffering and punitive damages are highly variable, the potential reliability benefits of CCG are great. As discussed above, the higher the variability of the award, the greater the potential benefits of CCG, and the greater our tolerance for bias. The potential for harm from bias arising from misalignment would be more concerning, for example, if a court were interested in applying prior-award information to reduce the already-low variability of an award for a certain type of economic damages. However, the high variability of awards for pain and suffering and punitive damages—sufficiently high to prompt drastic (and highly-biasing) measures such as damage caps—suggests that the benefits with respect to variability are likely to dominate any bias arising from misalignment (or any negative effects arising from high claim variability).

Second, the bias introduced by CCG methods is likely to be low. Statistically, the appropriateness of a set of “comparable” cases, relative to the present case, involves three major factors: 1) the “alignment” of the mean of the “correct” awards in the prior cases with the “correct” award in the present case, 2) the variability of the correct awards in the prior cases, and 3) the number of prior cases, or the “sample size.” In practice, a court
should be concerned with the alignment of the material facts of the prior cases with those of the present case, the substantive breadth of the prior cases, and the sample size.

Note that we cannot perfectly predict how a trier of fact will in fact incorporate guidance from prior awards. For purposes of this discussion, and for the guidance I provide in Section 5 below regarding the development of comparable-case information, I assume that the trier of fact will not act “irrationally” by affording more influence to highly-dispersed prior awards than to minimally-dispersed prior awards. It is not necessary for the trier of fact to behave perfectly rationally, but the foregoing assumption provides a good starting point for guidance. We “trust” the trier of fact to act approximately rationally in numerous legal contexts; and in developing guidance for the trier of fact, the court arguably should assume some degree of rationality, as it does in providing other forms of guidance.

Statistically (using fundamentals of “shrinkage estimation”), incorporating prior-award information will generally improve reliability, as long as the prior awards are not tightly bound (i.e., with low variability) around a mean that is far from the correct award in the present case— in practice, as long as the prior awards are not tightly clustered around an award that reflects facts materially dissimilar to the present case.

High prior-award variability itself is (absent extreme conditions) unlikely to harm the damages award, since the variability of awards for pain and suffering and punitive damages is high, and, as suggested by the discussion above, the weight afforded the prior-

---

69 See generally Efron & Morris, supra note 59; James & Stein, supra note 67.
award information (and, statistically, the mean of the prior awards in particular) is proportional to the inverse of the variability of the prior awards. This means that highly dispersed prior awards (e.g., prior awards that span a wide range of values) will have little influence on the award determination (in the model above, \( \hat{\alpha}_s \) will be approximately equal to \( \alpha \)).

Additionally, although uncomparable prior cases can cause significant bias, this concern fades when we consider the relatively remote conditions necessary for CCG to harm reliability, in the sense of increasing award error. In particular, for CCG to harm reliability, it would generally require a combination of significantly dissimilar cases and low prior-award variability—a combination that is highly unlikely if the prior cases are specifically selected to be comparable to the present case, and where the variability of the awards can often reflect (and perhaps should reflect) the court’s level of confidence regarding the comparability of the prior cases to the present case.\(^7\)

In short, we can expect that a reasonable method for identifying prior “comparable” cases will result in prior-award information that is likely to improve reliability.\(^7\) Of course, the more comparable the prior cases are to the present case, the better the guidance for the trier of fact, and the greater the reliability benefits. In the following section, I discuss in greater detail the development of comparable-case information.

\(^7\) See Section 5 for a discussion of methods for selecting comparable cases.

\(^7\) A formal derivation of the particular mean and variability conditions that would cause a reduction in reliability is beyond the scope of the current article.
5. Developing Comparable-Case Information

In this section, I discuss a number of considerations for establishing comparable-case (or “prior-award”) information. I address two areas of concern: 1) identifying a set of materially “comparable” cases, and 2) distilling comparable-case information from these cases for consideration by the trier of fact.

5.1 Identifying “Comparable” Cases

To use CCG, a court must identify a set of comparable cases. As suggested above, a court should be concerned with three factors: the alignment of the prior cases with the subject case, the breadth of the prior cases, and the number of prior cases. The suitability of the cases and expected benefits of the guidance depend on the courts care in balancing these factors. For example, prior awards with lower breadth generally produce, in a sense, more “specific,” and therefore more influential, guidance; but lower breadth also results in a smaller sample size, whereas a larger sample provides more information and therefore better guidance. Additionally, the alignment of the prior cases affects the bias introduced by the prior awards, and the breadth of the prior cases, which affects the weight or influence of the prior awards, may reflect the court’s confidence in the alignment of the cases identified.

In terms of alignment and breadth, the most significant factor affecting a court’s identification of prior cases is likely its understanding of the subject case and which facts and issues are fundamental to an appropriate comparison.

Consider, for example, Judge Posner’s analysis in *Jutzi-Johnson*. That case involved allegations that a jail was negligent in failing to identify a prisoner as a suicide risk and take
precautionary measures, resulting in the prisoner hanging himself. Commenting on the parties’ identification of comparison cases, Judge Posner explained:

The plaintiff cites three cases in which damages for pain and suffering ranging from $600,000 to $1 million were awarded, but in each one the pain and suffering continued for hours, not minutes. The defendant confined its search for comparable cases to other prison suicide cases, implying that prisoners experience pain and suffering differently from other persons, so that it makes more sense to compare Johnson’s pain and suffering to that of a prisoner who suffered a toothache than to that of a free person who was strangled, and concluding absurdly that any award for pain and suffering in this case that exceeded $5,000 would be excessive. The parties should have looked at awards in other cases involving asphyxiation, for example cases of drowning, which are numerous . . . . Had they done so, they would have come up with an award in the range of $15,000 to $150,000.72

Judge Posner thus disagreed with the alignment of the plaintiff’s cases, which involved pain and suffering that continued for hours rather than minutes. He also disagreed with the alignment, as well as the breadth, of the defendant’s cases, which were “confined” to “other prison suicide cases”—as though “prisoners experience pain and suffering differently from other persons”—leading to “absurd” conclusions regarding a suitable award.73 Finally, Judge Posner displayed concern for sample size, suggesting that widening the breadth of the prior cases to those involving asphyxiation would have resulted in “numerous” prior cases for comparison.74 Sample size is important because prior awards are subject to judgment variability: if a court selects only a single case for comparison, then, even if the court is extremely confident that the prior case is indeed comparable, it would remain unclear whether the award in that case is due to the specific facts of the case or to

---

72 Jutzi-Johnson v. United States, 263 F.3d 753, 760-61 (7th Cir. 2001).
73 Id.
74 Id.
randomness. A court may be justified in sacrificing some degree of “comparability” for purposes of increasing the sample size. Of course, in practice, a court may determine that the material facts of a case are too unique for comparison, and may accordingly forego any application of CCG.

With these factors in mind, the court should arrive at a process for selecting a suitable set of prior cases. Selection methods should remain flexible so that courts may cater such methods to the unique circumstances of the case at hand. Initially, courts can obtain guidance from comparative approaches used regularly to determine damages in various civil contexts.75

Thus, consider a simple procedure in which the court asks each party to identify a set of “comparable” cases. Such a procedure is likely to result in a set of extreme (although perhaps unbiased) cases, since each party selects cases to maximize its own interests.76 On the other hand, a method by which the court, rather than the parties, selects a set of “comparable” cases is likely to produce less variability, but is susceptible to a number of drawbacks, including the potential for judicial bias and undue burden on the court. Alternatively, the court may direct the parties to agree on a set of prior cases; but this may be difficult and inefficient without additional structure from the court.77

75 See infra Section 6. Courts may also obtain guidance from the class action context, where courts sometimes select a set of “representative” or “bellwether” cases for settlement or litigation purposes. See generally Bavli, supra note 50.

76 See generally Loren H. Brown et al., Bellwether Trial Selection in Multi-District Litigation: Empirical Evidence in Favor of Random Selection, 47 AKRON L. REV. 663, 674-75 (2014).

77 See generally id. at 675.
Although the circumstances of a particular case may call for one of the methods above, a particularly useful approach is one in which the court selects a final set of cases from a pool identified by the litigants. Variations of this method have been used in various civil contexts.\textsuperscript{78} It allows for significant flexibility, and benefits from both a litigant-based selection process and the relative objectivity of the judge.

Finally, it is possible to include (implicitly) the trier of fact in the selection process by providing it with qualitative information regarding the set of comparable cases, thus allowing it to weight the prior-award information based, in part, on its assessment regarding the alignment of the prior cases. I discuss this possibility further in the following subsection.

\subsection*{5.2 Distilling Comparable-Case Information}

Once the court has selected a suitable set of prior cases, it must decide what information should be considered by the trier of fact as guidance. A study by Professor Michael Saks et al. examined the effects of various forms of jury guidance on amounts awarded for pain and suffering in personal injury cases.\textsuperscript{79} They found that, while damage caps performed poorly in terms of variability and distortion of award magnitude, various forms of distributional information—in particular, an “award interval,” an “award average-plus-interval,” and “award examples”—were effective in reducing variability while

\textsuperscript{78} See infra Section 6; see also Brown et al., supra note 76, at 672-76 (describing cases employing such procedures in the bellwether context). See generally Alexandra D. Lahav, The Case for “Trial by Formula”, 90 Texas L. Rev. 571 (2012).

\textsuperscript{79} See Saks et al., supra note 15.
distorting award magnitude minimally or not at all.\textsuperscript{80} These results support the model described in Section 4. In particular, in line with the analysis above, prior-award information is expected to cause a reduction in variability and minimal (if any) distortion of award magnitude where the distribution of prior-awards is similar to the distribution that would be obtained in the absence of prior-award information. Indeed, Saks et al. obtained the values of the guidance provided to survey participants from the results of their pilot study. Specifically, they used the median of the pilot study for the award average, the 10th and 90th percentiles for the award interval, the 10th, 50th, and 90th percentiles for the award average-plus-interval, and the 10th, 35th, 65th, and 90th percentiles for the award examples.\textsuperscript{81}

The discussion in Section 4 suggests that courts should choose information forms that provide jurors with information regarding the “center,” as well as the “spread,” or variability, of the distribution of the prior awards.\textsuperscript{82} The award interval, award average-plus-interval, and award examples have the advantage, relative to providing, for example, only the award average, that jurors are provided information regarding the variability of the prior awards. This enables them to weight the prior-award information appropriately. The award examples form is particularly appealing, because it provides jurors with the

\textsuperscript{80} See id. at 249-55. Saks et al. observed mixed results for the “award average” form. See id. at 253-55.

\textsuperscript{81} Id. at 248. The authors’ results for the cap condition, which caused a distortion in the size, or magnitude, of the award and mixed effects on variability (depending on the severity category) can be interpreted as supportive of the model herein, but may also have implications for the effects of prior-award information.

\textsuperscript{82} It is possible that, empirically, jurors will not in fact weight the prior-award information according to the relative variability of such information; but without significant evidence to the contrary, courts should arguably provide jurors with relevant variability information.
most information, relative to the other information forms, regarding prior-award variability.

Further, in addition to numerical prior-award information, it may be beneficial for the court to allow the trier of fact to hear qualitative summary information regarding the prior cases. Providing the trier of fact with such information would strengthen its role in the possibly fact-intensive identification process, and would mitigate the effects of potential bias from misalignment. As suggested above, providing the trier of fact with qualitative case information implicitly includes it in the case-selection process by enabling it to weight the prior-award information based, in part, on its assessment of the comparability of the prior cases.83

If the court allows the trier of fact to hear qualitative case information, it must determine an appropriate form for such information—for example, a brief description regarding the set of cases as a whole, or a set of short factual summaries regarding each individual case.

As described in Section 6 below, there is substantial precedent for allowing a jury to hear comparable-case information in the form of evidence introduced as expert testimony. A jury may thereby “consider . . . arguments about the effect of [purported differences] on the validity of [the] comparison and . . . adjust its damage award accordingly.”84

---

83 The court should balance the benefits of allowing qualitative case information with the costs (including attorneys’ fees, delays, etc.) of allowing such information.

84 Syufy Enterprises v. American Multicinema, Inc., 793 F.2d 990, 1003 (9th Cir. 1986).
6. Conclusion

In the previous sections, I addressed three major objections to CCG methods. First, I explained that CCG can reduce unpredictability and improve the reliability of awards—notwithstanding the judgement variability associated with the prior awards—by allowing for the sharing of information across cases. Second, I discussed how information regarding awards in comparable cases can be understood, not as an alternative to assessing damages for pain and suffering or an appropriate punitive damages award, but as fundamental to assessing such values. Finally, I explained that it is not necessary for the prior cases to be perfectly comparable to the subject case, and that, assuming a reasonable method for identifying “comparable” cases, it is unlikely—particularly for awards for pain and suffering and punitive damages—that they will diminish, rather than improve, reliability.

Moreover, it is important to understand that using comparable-case information as guidance in determining awards for pain and suffering and punitive damages is not a remote concept that requires dramatic change to current practices. Indeed, courts have recognized the importance of comparable-case information as guidance in determining awards in a range of contexts.

First, there is substantial precedent from the judicial-review context for allowing information regarding prior awards to influence present awards and, in particular, for the use of comparable-case information as guidance in determining an appropriate award or range of awards. Indeed, courts have described the comparison of awards to prior awards for similar injuries as “[a] mainstay of the excessiveness determination,” noting that “[t]his use of comparison is a recognition that the evaluation of emotional damages is not readily
susceptible to ‘rational analysis.’”85 For example, courts in New York have held that, in determining whether an award is excessive, a reviewing court must consider awards in comparable cases86—that the “determination of whether a compensatory damages award is excessive should not be conducted in a vacuum, but instead should include consideration of the amounts awarded in other, comparable cases.”87 The U.S. Supreme Court has itself indicated, in the context of determining whether a particular punitive damages award exceeded the amount permitted by the Due Process Clause of the Constitution, that “[c]omparing the punitive damages award and the civil or criminal penalties that could be imposed for comparable misconduct provides a[n] . . . indicium of excessiveness.”88

Second, there is substantial precedent from the bench-trial context for CCG—and specifically, for consideration of comparable-case information by the trier of fact for guidance in determining awards for pain and suffering and punitive damages. In fact,

85 Salinas v. O’Neill, 286 F.3d 827, 830-31 (5th Cir. 2002) (internal citations omitted).


87 DiSorbo v. Hoy, 343 F.3d 172, 183 (2d Cir. 2003) (internal quotation marks and citations omitted). See Geressy v. Digital Equip. Corp., 980 F. Supp. 640, 656 (E.D.N.Y. 1997) aff’d in part sub nom. Madden v. Digital Equip. Corp., 152 F.3d 919 (2d Cir. 1998) (“CPLR 5501(c)’s conception of reasonable compensation cannot exist in a vacuum. There needs to be some point of reference. With economic damages, the court may rely on traditional methods of economic analysis. As for the non-economic pain and suffering award, the reviewing court must begin by identifying some group of similar cases to serve as a referent. This task is difficult, particularly in cases exploring relatively new types of injuries and claims such as those in the instant case involving [repetitive stress injury (RSI)] claims against a keyboard manufacturer. Cases with similar causal agents, similarly-named diagnoses, or similar reductions in quality of life might serve as benchmarks.”).

88 See BMW of N. Am., Inc. v. Gore, 517 U.S. 559, 583 (1996). See also Degorski v. Wilson, No. 04 CV 3367, 2014 WL 3511220, at *1 (N.D. Ill. July 16, 2014) (“In assessing whether a punitive damage award is constitutionally appropriate, the Supreme Court has directed courts to focus their evaluation on three guideposts: (1) the reprehensibility of the defendant’s conduct; (2) the relationship between the amount of the punitive damages awarded and the harm or potential harm suffered by the Plaintiff; and (3) the difference between the punitive damages award and the civil penalties authorized or imposed in comparable cases.”).
consideration of comparable-case information is often expected. For example, in *Jutzi-Johnson*, Judge Posner remarked that 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."\(^8^9\) He advised:

> To minimize the arbitrary variance in awards bound to result from such a throw-up-the-hands approach, the trier of fact should . . . be informed of the amounts of pain and suffering damages awarded in similar cases. 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. We make such comparisons routinely in reviewing pain and suffering awards, as do other courts. It would be a wise practice to follow at the trial level as well.\(^9^0\)

Third, there is precedent from various civil contexts for asking jurors to make comparisons for purposes of determining damages. The comparisons often relate to goods or services for which there is an economic market, and the comparisons are undertaken for purposes of providing information regarding the market, which, in turn, provides information regarding the appropriate damages award.

For example, in the context of determining “just compensation” in eminent domain litigation, it is common for a court to ask the jury to undertake a “comparable sales”

---

\(^{8^9}\) *Jutzi-Johnson v. United States*, 263 F.3d 753, 759 (7th Cir. 2001).

\(^{9^0}\) *Id.; see also Arpin v. United States*, 521 F.3d 769, 776-77 (7th Cir. 2008) (explaining that the lower court “should have considered awards in similar cases, both in Illinois and elsewhere,” and that “[c]ourts may be able to derive guidance for calculating damages for loss of consortium from the approach that the Supreme Court has taken in recent years to the related question of assessing the constitutionality of punitive damages”); *Zurba v. United States*, 247 F. Supp. 2d 951, 961 (N.D. Ill. 2001) aff’d, 318 F.3d 736 (7th Cir. 2003) (considering awards in “comparable” cases); *Elk v. United States*, 87 Fed. Cl. 70, 97 (2009) (“[v]arious cases, including those arising under the FTCA, suggest that judges faced with the prospect of determining pain and suffering awards should look to the awards in similar cases”); *Cochran v. A/H Battery Associates*, 909 F. Supp. 911, 917 (S.D.N.Y. 1995) (“courts have long reviewed damage awards in other similar cases to determine the scope of reasonableness in a particular case”).
analysis, for which the jury hears testimony regarding the selling price of comparable properties.\textsuperscript{91} In such cases, the court intends to provide the jury with information regarding the market to which the property at issue belongs, and thereby facilitate the jury’s assessment of just compensation and thus an appropriate compensatory damages award. Indeed, comparable sales evidence is often viewed as the “best evidence” of market value.\textsuperscript{92} Similarly, damage awards for breaches of contract for the sale of goods or services are often computed by looking at the market price of comparable goods or services.\textsuperscript{93} And courts also rely on the jury to make comparisons in arriving at damages calculations in the antitrust context.\textsuperscript{94}

\textsuperscript{91} See United States v. 1,129.75 Acres of Land, More or Less, in Cross & Poinsett Cys., 473 F.2d 996, 998 (8th Cir. 1973) (citing cases, and holding that, “[g]enerally, evidence of sales of comparable property is persuasive evidence of market value, either as direct proof or in support of a witness’s opinion”).

\textsuperscript{92} See United States v. 24.48 Acres of Land, 812 F.2d 216, 218 (5th Cir. 1987) (“First, we reiterate the principle that the best evidence of market value is comparable sales”) (citing United States v. 8.41 Acres, 680 F.2d 388, 395 (5th Cir. 1982)). See also John J. Dvorske & Ann K. Wooster, Condemnation of Property, Comparable Sales, 7 FED. PROC., L. ED. §14:90 (December 2015).

\textsuperscript{93} U.C.C. §2-713 (“the measure of damages for non-delivery or repudiation by the seller [of goods] is the difference between the market price at the time when the buyer learned of the breach and the contract price together with any incidental and consequential damages provided in th[e] Article (Section 2-715), but less expenses saved in consequence of the seller’s breach”). See also Gulf Power Co. v. Coalsales II, LLC, 522 F. Appx. 699, 707 (11th Cir. 2013) (holding, in a case involving a breach of a contract for the purchase and sale of coal, that “the appropriate measure of damages would be the market remedy provided for in §672.713 [of the U.C.C.], in which damages are calculated by comparing the contract price to the fair market value of comparable goods on the open market”).

\textsuperscript{94} For example, in Syufy Enterprises v. American Multicinema, Inc., 793 F.2d 990 (9th Cir. 1986), counterclaimant AMC based its damages calculation on comparisons with purportedly comparable markets. It presented such comparisons to the jury, which ultimately rendered an award accordingly. See id. at 1002-03. The Ninth Circuit rejected the argument that damages may not be computed based on comparable markets. See id. The court then rejected the argument that the comparison markets were not comparable to the market at issue and thereby inappropriately inflating AMC’s damages calculation. See id. at 1003. The court held that “[c]omparability is a question of fact” and that “[i]t was for the jury to consider [the opposing party’s] arguments about the effect of the [claimed differences] on the validity of comparison and to adjust its damage award accordingly. There was sufficient evidence to allow a jury to render a damage award for AMC on the basis of the comparison . . . .” Id.
A detailed examination of the legal foundations of CCG is beyond the scope of this article. The legal context above is intended simply to show that there is a long history, across various contexts, of using comparable-case information as guidance in determining damages. Indeed, courts have applied such methods notwithstanding the objections addressed in this article. For example, consider the objection that courts will compound the arbitrariness of awards by obtaining guidance from prior awards that presumably suffer from the same arbitrariness that courts hope to address. This objection applies similarly to comparable-case information considered in the bench-trial context, but bench-trial judges nevertheless consider comparable-case information. Additionally, the objection that methods involving the use of prior awards “fail to address the fundamental issue of how one should initially assess the value of pain-and-suffering damages” or arrive at an appropriate punitive damages award applies similarly to the use of comparable-case information in the judicial-review and bench-trial contexts; but such information is still used in these contexts. Finally, the objection that the validity of methods involving comparable-case information relies heavily on the presumption that the prior cases identified are indeed materially comparable to the present case applies similarly to all of the contexts discussed above, but such methods are nevertheless used.

My aim in this article was to address the major objections to CCG methods by explaining simply, but formally, what many courts and commentators have recognized

---

95 Avraham, supra note 19, at 104.

96 The objection applies directly to the judicial-review and bench-trial contexts and analogously to the market-comparisons context.
implicitly: that CCG is effective in reducing unpredictability and improving the reliability of awards for pain and suffering and punitive damages by allowing for the sharing of information across cases.
IV. Shrinkage Estimation in the Adjudication of Civil Damage Claims*

(Coauthor: Yang Chen)

1. Introduction

A legal proceeding can be understood as a procedure for generating an outcome that serves as an estimate of the “correct” outcome associated with a legal claim. In this sense, a criterion for measuring the strength of a legal procedure is reliability: the degree to which the procedure can be expected to generate “accurate” outcomes, outcomes that are close in proximity to the “correct” outcome.¹ In two recent articles written by one of the authors, it is argued that certain claim aggregation methods, methods in which the outcome of a claim is based not only on the characteristics of the claim itself, but also on the outcomes of other claims, can improve outcomes generally. The first article examines the conditions under which sampling procedures can improve accuracy in the class action context,² while the second article examines the use of comparable-case guidance (CCG), or “prior-award information”—information regarding awards in comparable cases as guidance for determining damage awards—to improve accuracy in the individual-claim context.³

---

¹ Hillel J. Bavli & Yang Chen, "Shrinkage Estimation in the Adjudication of Civil Damage Claims," REV. OF L. & ECON. (forthcoming, 2017). In writing this paper, the authors benefitted from the guidance of Donald B. Rubin and Jun S. Liu.

² See Aggregating for Accuracy.

Although sampling and comparable-case guidance are quite distinct in practice, and arise in different contexts, the underlying mechanisms by which they affect accuracy are similar.

Sampling procedures involve adjudicating a proportion of claims in a class action (the claims in the “sample group”) and extrapolating damage awards for the remaining claims (the claims in the “extrapolation group”). CCG methods involve incorporating information regarding awards in prior comparable cases in the adjudication of a damages award in the subject case. Sampling allows for the sharing of information across claims in a class, whereas CCG allows for the sharing of information across individual claims—but both methods aggregate and use information regarding awards in comparable claims to influence awards of other claims.

In the current article, we examine a third, but closely related—and in a sense unifying—form of claim aggregation that integrates such influence explicitly. This form of claim aggregation is based on a statistical device called “shrinkage estimation” (or “shrinkage”), which is used in statistics to aggregate information and thereby improve estimation. Specifically, it involves adjusting an estimate of some value to account for information derived from the population of units from which that value is drawn. CCG, which uses information regarding comparable claims to influence the subject claim, can be understood as a form of shrinkage. Similarly, sampling constitutes a special case of shrinkage, where the population of units is the class of claims, and the influence of the

---

information derived from an individual claim entirely determines the award for each claim in the sample group, and the influence of the information derived from sampled claims (rather than an individual claim) entirely determines the award for each claim in the extrapolation group.

Our objectives in this article are to examine the conditions under which shrinkage can increase the accuracy of damage awards in the class action and individual-claim contexts, and to apply shrinkage to gain a deeper understanding of the benefits and limitations of the foregoing methods with respect to accuracy.

We begin in Section 2 by reviewing the sampling framework developed in *Aggregating for Accuracy*. In Section 3, we build on this framework to examine the benefits of shrinkage in the class action context, and to reexamine the benefits of sampling in light of shrinkage estimation. We consider alternative methodologies under various assumptions regarding cost and legal constraints. In Section 4, we examine conditions under which shrinkage can be used to increase the accuracy of damage awards in the individual-claim context. In particular, we consider shrinkage in the context of providing jurors with prior-award information as guidance in determining awards for pain and suffering and punitive damages, and we derive and illustrate the conditions under which such guidance can in fact improve accuracy. In Section 5, we conclude.

2. A Framework for Examining Sampling and Accuracy in Class Action Litigation

In this section, we summarize the framework developed in *Aggregating for Accuracy* and a number of central results related to the use of sampling to improve accuracy in class
action litigation—first in the context of a homogeneous class, and then in the context of a heterogeneous class.

### 2.1 Sampling in a Class of Homogeneous Claims

*Aggregating for Accuracy* develops a framework for examining the effect of sampling on accuracy in class action litigation. The article examines a procedure in which 1) a number of claims are sampled from a class of claims for individualized adjudication (and receipt of individualized damage awards), and 2) the mean of the awards adjudicated in the sample group is applied as the award for all remaining claims, the claims in the extrapolation group.

The article's analysis is intended to respond to arguments that such procedures increase efficiency (by allowing the claimants to proceed as a class rather than as individual claimants), but only at the cost of reducing reliability. It builds on assertions put forth by Professors Saks and Blanck in a 1992 Stanford Law Review article to argue that, under certain conditions, sampling and extrapolation methods can increase reliability by reducing error associated with judgment variability—uncertainty resulting from variability in the composition of a jury, the presentation of evidence, the selection of a judge, etc.\(^5\)

To illustrate, consider how replication may be used to reduce judgment variability:

[I]magine now a (costly) hypothetical procedure in which each and every claim [in a given class] were litigated ten times independently, and in which the outcome associated with each claim were computed by taking the average of the ten verdicts associated with that claim. That is, start by taking the first claim and litigate it before ten independent juries to obtain ten independent

---

verdicts. Then assign the average of the ten verdicts as the outcome of the first case. By applying this aggregated outcome, rather than any single verdict, we may reduce the error resulting from judgment variability to nearly nothing.\textsuperscript{6}

\textit{Aggregating for Accuracy} shows that sampling in the context of a class of homogeneous claims, through its implicit or explicit use of replication, reduces judgment variability and thereby to increases accuracy.\textsuperscript{7} In particular, the article concludes that given a class of $N$ homogeneous claims, and legal restrictions that can be described by “reductive sampling”—that is, conditions under which the court will not replace an individually adjudicated award with an award extrapolated from other claims—accuracy is maximized by randomly selecting a sample of $n^* = \sqrt{N}$ claims for individualized adjudication, assigning the individualized awards as the outcomes of the claims in the sample group, respectively, and then applying the sample mean of the sample-group awards as the outcome of all remaining $(N - \sqrt{N})2$ claims in the class.\textsuperscript{8}

To be more precise: Assume we have a class of $N$ homogeneous claims from which we sample $n$ claims. Let us define a “correct” outcome associated with a given claim as the average award that would emerge from repeated litigation under various circumstances, such as various jury combinations, various lawyers, judges, and presentations of evidence.\textsuperscript{9} The “correct” outcome can be defined using various measures of central tendency, such as the mean or median. Throughout the paper, we adopt the measure used in \textit{Aggregating for Accuracy} at § 4.

\begin{itemize}
\item \textsuperscript{6} See \textit{Aggregating for Accuracy} at 77.
\item \textsuperscript{7} See id at § 4; \textit{Justice Improved} at 815-19.
\item \textsuperscript{8} \textit{Aggregating for Accuracy} at § 4.
\item \textsuperscript{9} \textit{Aggregating for Accuracy} at § 3; \textit{Justice Improved} at 833-34.
\end{itemize}
Accuracy and Justice Improved—the mean. Thus, let \( \mu \) be the correct award in each of the \( N \) homogeneous claims—the mean of the awards over repeated litigation of any (or all, since the claims are homogeneous) of the \( N \) claims. And let \( X_i \) be a random variable defined by the adjudicated award in the \( i^{th} \) claim, for \( i = 1,2,3,\ldots,n \), where the \( X_i \) are independent and identically distributed (i.i.d.) with mean \( \mu \) and variance \( \sigma^2 \), and the mean of the \( X_i \) for \( i = 1,2,3,\ldots,n \), is defined as \( X_n \) which is distributed with mean \( \mu \) and variance \( \frac{\sigma^2}{n} \). That is:

\[
X_i \sim (\mu, \sigma^2) \text{ and } X_n \sim (\mu, \frac{\sigma^2}{n})
\]

Note, in Aggregating for Accuracy, it is assumed that the \( X_i \) are distributed normally; but this distributional assumption is not necessary for the results derived in that paper or in the current paper. We therefore drop the normality assumption.

Thus, using the sum of square residuals

\[
R = \sum_{i=1}^{n} (X_i - \mu)^2 + (N - n)(\overline{X}_n - \mu)^2
\]

as the criterion for measuring error associated with all \( N \) claims, Aggregating for Accuracy concludes that total error, or “risk”—defined as the expectation of \( R \) above—is minimized, and accuracy is maximized, not by litigating each claim individually (a procedure often viewed by courts and scholars as the ideal, with respect to accuracy), but rather by sampling and individually adjudicating

\[
N_{n^*} = \sqrt{N}
\]

\(^{10}\) Aggregating for Accuracy at § 4.
claims, and applying the sample mean \( \bar{X}_n \) as the outcome in all remaining \((N - \sqrt{N/2})\) claims, the claims in extrapolation group.\(^{11}\)

Thus, in the context of a homogeneous class of claims, sampling may improve accuracy as well as efficiency. But, as explained in *Aggregating for Accuracy*, homogeneity is not necessary for sampling to increase accuracy. First, a court may stratify a heterogeneous class into multiple relatively homogeneous subclasses. For example, the District Court for the Eastern District of Texas, in *Cimino v. Raymark*,\(^{12}\) used such a procedure when it divided a heterogeneous class of asbestos claims into five disease categories. Second, although homogeneity is helpful, it is not required—the error-reducing benefits of sampling apply even to a class of heterogeneous claims.

### 2.2 Sampling in a Class of Heterogeneous Claims

Homogeneity allows the court to realize maximal accuracy by sampling \( \sqrt{N} \) claims. As a class becomes more heterogeneous, the utility of the sampling is reduced, since the benefits of reducing judgment variability must now be balanced with the error introduced by applying a single point estimate—the sample mean—to a class of heterogeneous claims. That is, the benefit with respect to judgment variability must be balanced with the cost with respect to claim variability. Thus, under conditions of heterogeneous claims, the optimal sample size will fall between \( \sqrt{N} \) and \( N \).\(^{13}\) Simply stated, extrapolating outcomes of

\(^{11}\) *Id.*


\(^{13}\) *Aggregating for Accuracy* at § 4.
unsampled claims can increase accuracy so long as the heterogeneity of the claims is not too large.

*Aggregating for Accuracy* models heterogeneous awards as draws from a normal distribution with means $\mu_i$ and variance $\sigma^2$. That is, contrary to a homogeneous class, where all claims have the same correct award $\mu$, in a heterogeneous class, each claim $i$ has a correct award $\mu_i$. Thus, $X_i \sim (\mu_i, \sigma^2)$, independent, where $\mu_i \sim (\mu_0, \tau^2)$, i.i.d. Note, here again we relax the normality assumption used in *Aggregating for Accuracy*.

Thus, again using the sum of square residuals, we have

$$R = \sum_{i=1}^{n} (X_i - \mu_i)^2 \sum_{j=n+1}^{N} (\bar{X}_n - \mu_j)^2$$

as the criterion for measuring the error associated with all $N$ claims. *Aggregating for Accuracy* minimizes the expectation of this expression, and derives the optimal sample size under conditions of heterogeneous claims to be

$$n^* = \sqrt{N} \sqrt{\frac{\sigma^2 + \tau^2}{\sigma^2 - \tau^2}}.$$

Thus, when $\tau^2$, the claim variability, is dominated by $\sigma^2$, the judgment variability, $n^*$ falls between $\sqrt{N}$ and $N$. If the claim variability is greater than judgment variability ($\tau^2 \geq \sigma^2$), then $n^* = N$. And when the claim variability is zero ($\tau^2 = 0$), then $n^* = \sqrt{N}$, the result obtained for a homogeneous class.\(^\text{14}\)

\[^{14}\text{Id.}\]
3. **Shrinkage in the Class Action Context**

The best aggregation method, with respect to accuracy, should be considered in light of relevant legal and cost constraints. As indicated earlier, the framework and conclusions described above assume that a court may not replace an adjudicated award with an extrapolated award (an assumption referred to in *Aggregating for Accuracy* as “reductive sampling”). *Aggregating for Accuracy* argues that, while there is clear precedent for aggregation and extrapolation, replacing individually adjudicated awards with extrapolated awards is likely violative of the Constitution as well as rules and norms of civil procedure. In the absence of this constraint, however, other aggregation methods may be more beneficial with respect to accuracy. For example, assuming no legal or cost constraints, a court may adjudicate all claims individually and then replace all individual awards with extrapolated awards, such as with the mean of the individual awards.

In the current section, we relax the legal constraints presumed in *Aggregating for Accuracy* in order to consider the effect of shrinkage—which involves the replacement of an adjudicated award with one that is influenced by awards in comparable claims—on accuracy in the class action context. Relaxing these constraints is useful for at least two reasons. First, there may be contexts in which such procedures are permissible. For example, parties may opt for such procedures in settlement or mediation contexts. Second, examining the effects of shrinkage allows a more complete understanding of aggregation in light of relevant legal and cost constraints.

Thus, in the current section, we begin by showing that, in the context of a class of claims, shrinkage achieves greater accuracy than classical case-by-case adjudication. Our
point of comparison is case-by-case adjudication (as in *Aggregating for Accuracy*), rather than a typical representative action, because the former is often viewed as the ideal, with respect to reliability, and is used as the primary alternative to class certification if putative class representatives are unable to show that class treatment is appropriate. We then reexamine the sampling results derived in *Aggregating for Accuracy* using shrinkage, and show that relaxing the reductive sampling constraint and applying shrinkage leads to greater accuracy than even the sampling procedure examined in that paper.

### 3.1 Comparison to Individualized Adjudication

Our objective in this section is to show that, given a class of claims, replacing an adjudicated damages award with an award based on shrinkage increases accuracy, in expectation, for each adjudicated claim (and, therefore, in the aggregate, over all sampled claims).

As above, assume we have a class of $N$ claims and that each claim $i$ is associated with a correct award $\mu_i$ for $i = 1, 2, 3, ..., N$. Assume the $\mu_i$'s are equal (homogeneous) or that they arise from a common distribution with mean $\mu_0$ and variance $\tau^2$. Denote the awards of the $n$ sampled claims by $X_1, ..., X_n$ and assume that they are distributed around their correct awards $\mu_i$, respectively, with variance $\sigma^2$. Thus, $X_i \sim (\mu_i, \sigma^2)$, independent, where $\mu_i \sim (\mu_0, \tau^2)$, i.i.d., and $\sigma^2$ and $\tau^2$ are known and represent judgment variability and claim variability, respectively. (See *Aggregating for Accuracy* for a discussion regarding parameter estimation). As above, we do not rely on distributional assumptions. Our objective is to impute all of the missing outcomes $\{\mu_i\}_{i=1}^N$, which are not directly observable, using
estimated values $\{\hat{\mu}_i\}_{i=1}^N$, and to do so in a way that minimizes error, represented by the (standard) risk function,

$$R = E \sum_{i=1}^N (\hat{\mu}_i - \mu_i)^2 = E \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 + E \sum_{i=n+1}^N (\hat{\mu}_i - \mu_i)^2$$

We thus replace the classical estimator—an adjudicated award ($X_i$)—with a shrinkage estimator ($\hat{\mu}_i$), which combines the adjudicated award with additional information obtained from the other claims in the class, and thereby generates a more accurate award. We define the shrinkage estimator as:

$$\hat{\mu}_i = \frac{X_i/\sigma^2 + \mu_0/\tau^2}{1/\sigma^2 + 1/\tau^2}.$$

This estimator ($\hat{\mu}_i$) for the correct outcome in the $i^{th}$ claim thus differs from the classical estimator, the adjudicated award ($X_i$), in that $\hat{\mu}_i$ is a weighted average of the adjudicated award ($X_i$) and the mean ($\mu_0$) of the $\mu_i$, weighted by the inverse of the variability of each, respectively. Thus, $X_i$ and $\mu_i$ are weighted by their “information”—the inverse of the judgment variability ($\sigma^2$) and the inverse of the claim variability ($\tau^2$), respectively. Intuitively, smaller variability implies greater information, and therefore heavier weight.\(^{15}\)

Note, if we assume that the distributions of $X_i$ and $\mu_i$ are Gaussian (or “Normal”)—i.e., if we regard $X_i \sim N(\mu_i, \sigma^2)$ as the likelihood and $\mu_i \sim N(\mu_0, \tau^2)$ as the prior—we indeed have an explicit statistical justification for this estimator. It is the Bayes estimator, which is admissible. It is also the maximum likelihood estimator in the hierarchical model.

\(^{15}\) See Casella, supra note 4.
It follows that the risk of the shrinkage estimator is

\[ R_i^s := \mathbb{E} (\hat{\mu}_i - \mu_i)^2 = \mathbb{E} \left( \left( \frac{X_i}{\sigma^2} + \frac{\mu_0}{\tau^2} - \mu_i \right)^2 \right) = \left( \frac{1}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}, \tag{1} \]

which is smaller than \( R_i^c := \mathbb{E} (X_i - \mu)^2 = \sigma^2 \), the risk associated with the classical estimator \( X_i \) of \( \mu_i \).

Therefore, for each individual claim, assigning the shrinkage estimator as the damages award, on average, yields greater accuracy—that is, lower risk—as compared to the adjudicated award. Further, this result leads to the conclusion that applying the individualized shrinkage award \( \hat{\mu}_i^s \) to each claim in the class reduces total risk, since risk is reduced with respect to each individual claim.

Now, in general, we will not know the value of \( \mu_0 \). Thus, substituting it with the unbiased estimator \( \hat{\mu}_0 = \sum_{i=1}^{n} X_i / n \) for \( \mu_0 \) in the shrinkage estimator, we have the “empirical” shrinkage estimator

\[ \hat{\mu}_i = \frac{X_i / \sigma^2 + \hat{\mu}_0 / \tau^2}{1/\sigma^2 + 1/\tau^2}, \tag{2} \]

where \( \hat{\mu}_0 \sim N(\mu_0, \frac{\sigma^2 + \tau^2}{n}) \). Note that in the Gaussian case where the \( X_i \) are independent and have distribution \( N(\mu_i, \sigma^2) \), and \( \mu_i \) has prior distribution \( N(\mu_0, \tau^2) \), this estimator is called the empirical Bayes estimator,\(^\text{16}\) where we replace the hyperparameter \( \mu_0 \) with its maximum likelihood estimate. This empirical Bayes estimator, which converges to the Bayes
estimator as \( n \to \infty \), is asymptotically admissible. Therefore, statistically it is a justified estimator.

Thus, let us confirm that the risk associated with this estimator is less than the risk associated with the classical estimator, an adjudicated award. Letting

\[
A = \frac{-2}{\sigma^2 + \tau^2}
\]

we can write this “empirical” shrinkage estimator \( \hat{\mu}_i \) as \( \hat{\mu}_i = AX_i + (1 - A)\hat{\mu}_0 \). The risk of this estimator is

\[
R_{i}^{se} := \mathbb{E} (\hat{\mu}_i - \mu_i)^2 = \mathbb{E} \left[ A(X_i - \mu_i) + (1 - A)(\hat{\mu}_0 - \mu_i) \right]^2
= A^2 \sigma^2 + 2A(1 - A) \mathbb{E} (X_i - \mu_i)(\hat{\mu}_0 - \mu_i) + (1 - A)^2 \mathbb{E}(\hat{\mu}_0 - \mu_i)^2
\]

\[
= \left( \frac{1}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} + \frac{1}{n} \frac{\sigma^4}{\sigma^2 + \tau^2} \to R_{i}^{a} \ (n \to \infty)
\]

This risk is somewhat larger than \( R_{i}^{a} \) with \( R_{i}^{se} - R_{i}^{a} = n^{-1}\sigma^4/(\sigma^2 + \tau^2) > 0 \). Intuitively, since we have less information—we do not know the true value of \( \mu_0 \)—we have greater risk. However, as the sample size increases, this risk converges to the case in which we know the true \( \mu_0 \). That is, \( R_{i}^{se} \to R_{i}^{a} \) as \( n \to \infty \).

Importantly, however, although \( R_{i}^{se} > R_{i}^{a} \), \( R_{i}^{se} \) (the risk associated with the empirical shrinkage estimator) is significantly smaller than \( R_{i}^{a} \), the risk associated with the classical estimator (the adjudicated award). That is, \( R_{i}^{se} - R_{i}^{a} = (-1 + n^{-1})\sigma^4/(\sigma^2 + \tau^2) < 0 \). Thus, under the specified criterion, the empirical shrinkage estimator is also (in addition to the shrinkage estimator) a better estimator than an adjudicate award.
Note that if we do not have a way of estimating claim variability (a possibility discussed in *Aggregating for Accuracy*), we can nevertheless rely on Stein’s estimator,\(^{17}\) which does not depend on claim variability:

\[
\hat{\mu}_i^{JS} = \hat{\mu}_0 + (1 - \frac{n-3}{S}) (X_i - \hat{\mu}_0),
\]

where \(S = \sum_{i=1}^{n} (X_i - \hat{\mu}_0)^2\) is an unbiased estimator for \((n - 1)(\sigma^2 + \tau^2)\). In the Gaussian case (where \(X_i\) and \(\mu_i\) are Gaussian), \((n - 3)/S\) is an unbiased estimator for \((\sigma^2 + \tau^2)^{-1}\). The risk associated with Stein’s estimator can be derived under Gaussian assumptions;\(^{18}\) and it can be shown that this risk is less than the risk associated with the classical estimator as well. Indeed, Stein’s estimator will converge to the shrinkage estimator as \(n \to \infty\); and the risk associated with Stein’s estimator and the shrinkage estimator above will be asymptotically equal.

### 3.2 Sampling with Shrinkage Estimation

In this subsection, we reexamine the accuracy benefits of sampling, but now using the empirical shrinkage estimator \((\hat{\mu}_i)\), rather than an adjudicated award \((X_i)\), for the claims in the sample group, and, once again, extrapolating awards for the remaining claims (the claims in the extrapolation group) using the estimated global mean \(\hat{\mu}_0\).

Let us begin by deriving the risk associated with the empirical shrinkage estimator in the context of the sampling framework discussed above. The total risk for sampled and non-sampled claims is:


where $R_{i}^{se}$ (Equation 5) is the risk associated with a single sampled claim $i$ using the empirical shrinkage estimator (Equation 2).

We can thus see that the risk $R_S$ is a monotone decreasing function of $n$:

$$\frac{\partial R_S}{\partial n} = -\frac{\tau^4}{\tau^2 + \sigma^2} - \frac{N}{n^2}(\sigma^2 + \tau^2) < 0.$$ (6)

This means that the risk will continue to decrease as we increase the sample size $n$.

On the other hand, if we use the classical estimator, an adjudicated award $X_i$ to estimate $\mu_i$ we obtain the risk function,

$$R_C = \mathbb{E} \sum_{i=1}^{N} (X_i - \mu_i)^2 = \mathbb{E} \sum_{i=1}^{n} (X_i - \mu_i)^2 + \mathbb{E} \sum_{j=n+1}^{N} (\mu_0 - \mu_j)^2$$

$$= n\sigma^2 + (N-n)(\frac{\sigma^2}{n} + \frac{n+1}{n} \tau^2),$$

which, as derived in Aggregating for Accuracy and reviewed above, is minimized at

$$n^* = \sqrt{\frac{N \sigma^2 + \tau^2}{\sigma^2 - \tau^2}}.$$ (7)

The divergence of the risk of the empirical shrinkage estimator from that of the classical estimator is shown in Figure IV.1. Figure IV.1 plots the risk associated with a class of 1000 claims against the number of claims, $n$, sampled for individual adjudication. The figure shows the risk $R_C$ associated with the classical estimator (adjudicated awards) and the risk $R_S$ associated with the shrinkage estimator for sample sizes between 10 and 200.
As Figure IV.1 illustrates, $R_C$ is minimized pursuant to equation 7—where here, $n^*$ is approximately 33 (with $\sigma/\tau = 5$)—whereas $R_S$ is monotonically decreasing in $n$. As Figure IV.1 suggests, if we relax the reductive sampling constraint in *Aggregating for Accuracy*, and instead assume the permissibility of shrinkage estimation, the sample size that minimizes risk is found to be $N$, the total number of claims in the class.

(Note, if claim variability is unknown, the court can first sample a small number of claims for adjudication and use $S = \sum_{i=1}^{n} (X_i - \hat{\mu})^2$ to estimate $(n - 1)(\sigma^2 + \tau^2)$. In this case, $n^* = \sqrt{N(\frac{2(n-1)\sigma^2}{S} - 1)^{-\frac{1}{2}}}$. We can then obtain the asymptotic interval for this new estimated number of claims using the central limit theorem and the delta method. We can therefore obtain a range for the optimal sample size, and determine our choice based on some extraneous criteria.)

---

19 See generally *Aggregating for Accuracy* §§ 4-5.
Figure IV.1: Illustration of risk, plotted against the number of claims, $n$, sampled for adjudication, comparing the risk, $R_C$, associated with the classical estimator (adjudicated awards), and risk, $R_S$, associated with the shrinkage estimator, given a class of 1000 claims and sample sizes between 10 and 200.

Importantly, the foregoing results should not be interpreted as supporting an argument against sampling; the reliability benefits of sampling are clear and significant under the conditions, including the legal constraints, described in *Aggregating for Accuracy*. Rather, these results demonstrate the benefits of shrinkage. Indeed, shrinkage does not detract from the reliability benefits of sampling; rather, it sufficiently enhances the accuracy of individual estimates—*i.e.*, resulting from individualized adjudication and the replacement of individually adjudicated awards with shrinkage estimates—that, in a sense, it reduces the need for (*i.e.*, the relative benefits of) sampling.
Furthermore, it is significant that in cases in which judgment variability dominates claim variability ($\sigma^2 > \tau^2$), shrinkage estimation reduces risk at a high rate when $n < n^*$ and at a relatively low rate when $n > n^*$ (since, as $n$ approaches $N$, the second term on the right hand side of equation 6 approaches 0, and the first term dominates). Therefore, if the reductive sampling constraint is relaxed—and, in particular, if courts are willing to adjust individual adjudications to incorporate aggregate information—then, to balance concerns for accuracy with concerns for litigation costs, a court may construct a sampling-based trial plan pursuant to the results derived in *Aggregating for Accuracy*, but add the additional step of replacing the adjudicated awards of sampled claims with shrinkage estimates. That is, the court may sample

$$n^* = \sqrt{\frac{N \sigma^2 + \tau^2}{\sigma^2 - \tau^2}}$$

claims for individualized adjudication; assign individualized *shrinkage estimates* to the $n^*$ sample-group claims (based on their respective individualized awards as well as the mean of the sampled claims); and apply the mean of the awards in the sample group as the outcome of all non-sampled claims. This procedure is identical to the procedure derived in *Aggregating for Accuracy*, with one difference: here, the court would apply individualized shrinkage estimates, rather than individualized classical awards, as the outcomes of the claims in the sample group.

In concluding this section, we note that we do not intend to make normative statements regarding the appropriate use of shrinkage in litigation. For example, it is beyond the scope of this paper to address the constitutionality of shrinkage or related policy concerns. Instead, we simply intend to develop a more complete understanding of
aggregation, with respect to reliability, and to examine a number of key results regarding the reliability benefits of shrinkage.

Given our results above, choosing the aggregation approach to maximize accuracy depends on the applicable legal and cost constraints. For example, if there is no concern for the reductive sampling constraint or cost constraints, repeated adjudications of each claim in a class can lead to the greatest degree of accuracy. If there is concern for litigation costs but no concern for the reductive sampling constraint, sampling a proportion of claims pursuant to *Aggregating for Accuracy* and replacing the adjudicated awards in the sample group with shrinkage estimates (and still using the mean of the sample group to extrapolate awards for all remaining claims) may maximize accuracy, given the applicable cost constraints. Finally, if there is concern for legal constraints described by reductive sampling (whether or not there is also concern for litigation costs), the sampling procedure (without shrinkage) derived in *Aggregating for Accuracy* is the alternative that maximizes accuracy.

In the following section, we extend our discussion to circumstances outside of the class action context. In particular, we consider the reliability implications of shrinkage in the individual claim context, where instead of shrinking toward the estimated global mean of a defined class of claims, we shrink toward the estimated global mean of a population of awards in prior comparable cases.

4. **Shrinkage Estimation in the Individual-Claim Context**

In Section 3, we derived the potential accuracy benefits of replacing an adjudicated award with a shrinkage award in the context of a class of (homogeneous or heterogeneous)
claims. We showed that shrinkage may be beneficial even under conditions of high claim variability. The class action context provides a convenient starting point for the application of shrinkage, and aggregation procedures generally, since we are given a population of claims (presumably with relatively low claim variability) over which we are to aggregate. However, a heterogeneous class of claims bound together by some common thread should not be understood as far different, for purposes of shrinkage, from a population of prior claims that are similarly bound together by some common thread. Therefore, in the current section, we extend our discussion of shrinkage to the individual-claim context.

For purposes of this section, there are two major challenges to applying shrinkage in the individual-claim context. The first is the problem, highlighted above in our discussion of reductive sampling, that generally a court cannot legally replace an adjudicated damages award with an award extrapolated formulaically. The second is the problem of identifying a suitable set of prior comparable cases to be used for shrinkage.

It is beyond the scope of the current paper to examine the legality of replacing an adjudicated award with a shrinkage award. Rather, we couch our discussion of shrinkage here in the context of a recent article, written by one of the authors, examining procedures that aim to reduce the judgment variability of certain types of (particularly unpredictable) damage awards by providing a fact-finder, prior to its determination of the award, with information regarding awards in prior comparable cases. In particular, The Logic of CCG uses principles of shrinkage to explain the accuracy benefits of comparable-case guidance.

---

20 The Logic of CCG at § 1.
beyond the empirical result that such guidance reduces variability, and to address a number of major challenges to such methods. Providing a fact-finder with prior-award information may serve as an innovative way of using shrinkage to improve certain types of damage awards—after all, it enables a fact-finder to incorporate prior award information just as a shrinkage estimator would. The major difference is that a shrinkage estimator incorporates such information formulaically, whereas, with the methods examined in *The Logic of CCG*, we must rely on certain behavioral assumptions.

As described in that paper, there is substantial evidence that providing jurors with such information is effective in reducing judgment variability and influencing a damages award generally, but whether a juror can be said to act as a shrinkage estimator (*e.g.*, allowing prior-award information to influence the award in proportion to the inverse variability of such information) is currently being studied in a series of randomized experiments.

In the current section, we address the second challenge—the problem of identifying a set of prior comparable cases. Effectively, we assume that the factfinder in *The Logic of CCG* acts as a shrinkage estimator, and we examine the robustness of comparable-case guidance, with respect to accuracy, to variations in the set of prior comparable cases identified. In other words, our concern is this: assuming the fact-finder acts “rationally,” in the sense of optimally using prior-award information as a shrinkage estimator would, what choices of prior cases (from which prior-award information is obtained) will increase accuracy in expectation? For example, how “wrong” can a set of prior awards be before
prior-award information reduces accuracy? These questions are essential for determining policy surrounding such methods.

Although we do not know, at this point, whether juries act as predicted, applying shrinkage explicitly to this context enables an understanding of the potential benefits, and some of the potential risks, associated with the use of comparable-case guidance. At the very least, our analysis enables an understanding of the potential benefits of comparable-case guidance, and the robustness of these benefits to “incorrect” sets of prior awards.

4.1 Background: The Use of Comparable-Case Guidance to Reduce Award Variability

The problem addressed in The Logic of CCG is the unpredictability (i.e., judgement variability) of awards for pain and suffering and punitive damages—two types of awards for which the jury receives very little guidance in determining the award. The Supreme Court and lower courts have repeatedly expressed the importance of reducing the variability of such awards.21

The paper highlights the problems with relying on existing methods, such as additur and remittitur, tools used by courts to increase or decrease the amount of an award found to be inadequate or excessive. Although these tools can be useful, and can be used to incorporate prior-award information in various ways,22 and in conjunction with other methods, alone they address extreme awards only, rather than variability generally, and (in

---


22 See generally Joseph B. Kadane, Calculating Remittiturs, 8 LAW, PROBABILITY & RISK 125 (2009).
practice) they ordinarily address only excessive awards, rather than inadequate ones.\textsuperscript{23} Additionally, widespread use of such methods arguably replaces the fact-finder’s discretion with that of the court, raising constitutional and policy issues. Other methods, such as caps, arbitrarily draw cutoff points, leading to substantial bias and perverse outcomes.\textsuperscript{24}

As mentioned above, there is empirical evidence that providing the factfinder with information regarding awards in prior cases is effective in reducing variability. But such studies do not address the effect of prior-award information on \textit{accuracy}—that is, bias and variability.

\textit{The Logic of CCG} establishes a framework for examining the benefits and limitations of prior-award information in terms of accuracy; and it addresses a number of major challenges to the use of prior-award information to reduce variability, including the possibility of using award information from an “incorrect” set of cases.\textsuperscript{25} In the current section, in applying shrinkage estimation to the individual claim context, we analyze, and derive a number of significant results regarding, this latter challenge in particular.

\subsection*{4.2 Identifying Prior Cases}

As a preliminary matter, it is important to realize that there is no “correct” or “incorrect” set of prior cases. As discussed in \textit{The Logic of CCG}, the effect of prior-award information on accuracy depends on 1) the alignment of material facts and issues in the subject case to those in prior cases; 2) the substantive breadth of the prior cases; and 3) the

\begin{itemize}
\item \textsuperscript{23} \textit{The Logic of CCG} § 3.
\item \textsuperscript{24} \textit{Id}.
\item \textsuperscript{25} \textit{See Id.} at § 1.
\end{itemize}
number of prior cases—or the “sample size.” For example, the alignment of facts and issues affects the bias introduced by the prior awards—we would like for the average correct award in the prior cases to align with the correct award in the subject case. The breadth of the prior awards affects, e.g., the influence of the prior-award information on the subject award; but a set of prior cases that contains only identical, or almost identical, material facts and issues may result in a sample size of one or two, or even zero, prior awards. Thus, in identifying a set of prior cases, a court must balance its interests in maintaining a reasonable sample size, a reasonable breadth, and cases that involve facts and issues that are relatively aligned with those in the subject case.26

Consider the example of the Seventh Circuit case, Jutzi-Johnson v. United States,27 described in The Logic of CCG, involving an award for pain and suffering arising from circumstances in which a jail inmate committed suicide by hanging, due to a failure of the jail to supervise him appropriately. A court considering prior awards (as Judge Posner of the Seventh Circuit did) might consider whether to use only cases involving inmates who hung themselves, people who hung themselves from the general population, people who committed suicide from the general population, people who suffered from asphyxiation (e.g., drowning) from the general population, etc.28 If the court restricts its consideration to cases involving inmates who hung themselves, it may obtain only a tiny sample size; if the

---

26 Id. at §§ 4-5.
27 263 F.3d 753 (7th Cir. 2001).
28 See id. at 760-61.
court uses a wider breadth of cases, the prior awards will have less influence, and higher
risk of introducing bias.\footnote{The Logic of CCG at § 5.}

For a more detailed discussion of these considerations, see The Logic of CCG. The
important point, for purposes of the current analysis, is that there are tradeoffs between
breadth, alignment, and sample size—combinations of which correspond to various levels
of bias and variance, and therefore accuracy. The purpose of the current analysis is to
examine these effects to understand the conditions under which prior-award information
increases accuracy.

Thus, assume that we have an individual claim that receives award \( Y \). \( Y \) is centered
at the correct award \( \mu_y \) with judgement variability \( \sigma_y^2 \) (assumed to be known). The correct
award \( \mu_y \), is centered at \( \lambda_0 \) with variance \( \eta^2_0 \). That is, the correct award can be understood as
a single award from a distribution containing a population of awards from comparable
claims. The correct awards of the comparable claims, as well as that of the subject claim,
are distributed around a global mean \( \lambda_0 \) with variability \( \eta^2_0 \). Thus,

\[
Y \sim (\mu_y, \sigma_y^2), \quad \mu_y \sim (\lambda_0, \eta^2_0).
\]

To be clear, in statistical terms, by “comparable” claims or cases, we mean to suggest
that their awards somehow arise from the same global mean.

Now, if we know the global mean \( (\lambda_0) \) and variability \( (\eta^2_0) \) associated with this
population, the shrinkage estimator is

\[
\hat{\mu}_y = \frac{Y/\sigma_y^2 + \lambda_0/\eta^2_0}{1/\sigma_y^2 + 1/\eta^2_0}.
\]
Similar to Equation 1, we know that the risk of this estimator is \((\sigma^{-2} + \eta_0^{-2})^{-1}\), which is smaller than \(\sigma^2\), the risk associated with the classical estimator \(Y\) to estimate \(\mu_y\). Notice that this estimator requires us to know the values of \(\lambda_0\) and \(\eta_0\). A more realistic scenario is one in which we do not know these values, and instead need to estimate them based on prior awards. Assume that the identified set of prior cases involves awards \(X_1, \ldots, X_N\) centered at the correct awards \(\mu_1, \ldots, \mu_N\) with judgement variability \(\sigma^2\) (assumed to be known). That is,

\[
X_i \sim (\mu_i, \sigma^2), \quad \mu_i \sim (\lambda_0, \eta_0^2); \quad i = 1, \ldots, N.
\]

We can then use the following unbiased estimators to estimate \(\mu_0\) and \(\eta_0^2\):

\[
\hat{\mu}_0 = \frac{\sum_{i=1}^{N} X_i}{N}, \quad \hat{\eta}_0^2 = \frac{\sum_{i=1}^{N} (X_i - \hat{\mu}_0)^2}{N - 1} - \sigma^2.
\]

We therefore use the “empirical” shrinkage estimator,

\[
\hat{\mu}_y = \frac{Y/\sigma_y^2 + \lambda_0/\hat{\eta}_0^2}{1/\sigma_y^2 + 1/\hat{\eta}_0^2},
\]

which converges to the (non-empirical) shrinkage estimator as \(N \to \infty\). As in Section 3, replacing \(\lambda_0\) and \(\eta_0\) with unbiased estimators introduces some amount of uncertainty into the estimator and thus introduces some risk. However, as \(N \to \infty\), the risk converges to that of the (non-empirical) shrinkage estimator. Therefore, in terms of risk, we benefit from a sufficiently large \(N\), the number of prior cases. Thus, if the set of prior cases is too small—in the sense that the risk resulting from the estimated values of \(\lambda_0\) and \(\eta_0\) dominates the benefits of the empirical shrinkage estimator—a court can increase sample size by expanding the substantive breadth of the prior cases.

In the following subsection, we discuss the breadth and alignment of prior cases. Although, for the reason explained above, sample size is an important consideration in
determining breadth, the accuracy benefits of shrinkage are fairly robust to low sample size. Specifically, because the shrinkage estimator is influenced by sample size only through the estimation of the mean and variance of the distribution of prior awards, and because these quantities can be estimated reasonably well with a small sample size, a sample of 5 to 15 often suffices. Perhaps most importantly, as discussed in detail below, the accuracy benefits of shrinkage are quite robust to misalignment; therefore, even if the estimation turns out to be relatively poor, shrinkage can generally be expected to improve accuracy. This is especially true, in light of the remittitur device and appellate process, by which the courts moderate extreme outliers. For these reasons, although we remain conscious of sample size, for simplicity, it does not play a central role in our examples and illustrations below.

4.3 Breadth and Alignment of Prior Cases

We are interested in examining two considerations: breadth and alignment. To illustrate, consider the case of Jutzi-Johnson, discussed above. Judge Posner disagreed with the prior cases identified by both the plaintiff and the defendant. He explained that “[t]he plaintiff cites three cases in which damages for pain and suffering ranging from $600,000 to $1 million were awarded, but in each one the pain and suffering continued for hours, not minutes”; and that “[t]he defendant confined its search for comparable cases to other prison suicide cases, implying that prisoners experience pain and suffering differently from other persons, so that it makes more sense to compare Johnson’s pain and suffering to that of a prisoner who suffered a toothache than to that of a free person who was strangled, and concluding absurdly that any award for pain and suffering in this case that exceeded $5,000
would be excessive.” Judge Posner ultimately concluded that “[t]he parties should have looked at awards in other cases involving asphyxiation, for example cases of drowning, which are numerous.”

In the language of the current section, Judge Posner disagreed with the alignment of the plaintiff’s cases, implying that awards corresponding to cases involving hours, rather than minutes, of pain and suffering would be inappropriately high. He disagreed with the breadth (and the alignment) of the defendant’s cases, suggesting that a set of cases involving the pain and suffering of inmates, rather than the general population, is too narrow; and that the defendant’s focus on inmates led to alignment issues that resulted in “absurd” conclusions. Additionally, Judge Posner seems to suggest that the sets of cases identified by the parties suffered from small sample sizes as well, indicating that broadening the prior cases to include those involving asphyxiation in the general population would lead to numerous cases.

Thus, consider a case that receives an award $Y$, which is “drawn from” a distribution centered at the correct award $\mu_y$ with judgement variability $\sigma_y^2$ (known), where $\mu_y$ is centered at $\lambda_0$ with variance $\eta^2$. That is,

$$ Y \sim (\mu_y, \sigma_y^2), \quad \mu_y \sim (\lambda_0, \eta^2). $$

Assume that the court identifies a set of prior cases involving awards $X_1, \ldots, X_N$ with correct outcomes $\mu_1, \ldots, \mu_N$ centered at $\mu_0$ with variance $\tau^2$. Thus,

---

30 Jutzi-Johnson, 263 F.3d at 760.
31 Id.
\[ X_i \sim (\mu_i, \sigma^2_i), \mu_i \sim (\mu_0, \tau^2); \quad i = 1, \ldots, N. \]

Essentially this means \( X_i \sim (\mu_0, \psi^2) \) for \( 1 \leq i \leq N \), where \( \psi^2 = \sigma^2 + \tau^2 \). Now, consider an estimator of the form

\[
\hat{\mu}_y = \frac{Y}{\sigma_y^2 + \psi^2},
\]

(8)
The which we approximate by plugging in unbiased estimators

\[
\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}, \quad S = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1},
\]

for the unknown parameters \( \mu_0 \) and \( \psi^2 = \sigma^2 + \tau^2 \), respectively, to obtain:

\[
\hat{\mu}_y^0 = \frac{Y}{\sigma_y^2 + \psi^2},
\]

(9)
Again, as \( N \to \infty \), \( \hat{\mu}_y^0 \to \hat{\mu}_x \). Thus, the risk of this shrinkage estimator \( \hat{\mu} \) is

\[
R_y = \mathbb{E}(\hat{\mu}_y - \mu_j)^2 = \mathbb{E}\left(\frac{Y + \mu_0}{\sigma_y^2 + \psi^2} - \mu_j\right)^2 = \mathbb{E}\left(\frac{Y - \mu_y}{\sigma_y^2 + \psi^2} + \frac{\mu_0 - \mu_y}{\psi^2}\right)^2
\]

\[
= \left(\frac{1}{\sigma_y^2} + \frac{1}{\psi^2}\right) + \frac{(\mu_0 - \mu_j)^2 + \eta_0^2}{\psi^4},
\]

which is smaller than \( \sigma_y^2 \) when

\[
(\mu_0 - \lambda_0)^2 + \eta_0^2 < 2\psi^2 + \sigma_y^2.
\]

(10)
Let us consider the meaning of this condition and then examine a number of numerical examples to gain a deeper understanding of the conditions necessary for increasing accuracy. On the right-hand side of Equation (10), \( \psi^2 = \sigma^2 + \tau^2 \) is the total variance (the sum of claim variability and judgment variability) of the prior awards; and \( \sigma_y^2 \) is the judgment variability of the subject case. On the left-hand side of Equation (10), \( (\mu_0 - \lambda_0)^2 \) is the square of the misalignment (i.e., the difference) between the expected correct
award in the subject case and the mean of the prior awards; and $\eta^2_0$ can be thought of as the claim variability of the hypothetical population to which the subject case belongs. For simplicity, we can set $\eta^2_0 = 0$ and view the case as its own population. The condition would be satisfied, for example, if (1) the judgment variability of the subject award is greater than the hypothetical claim variability (e.g., where $\eta^2_0 = 0$), and (2) the breadth of the prior awards is greater than their misalignment. Of course, the effects of one can be offset by the effects of the other. For example, the effects of extreme misalignment can be offset by the effects of extreme judgment variability.

Further, in general, the greater the dispersion of the prior awards, the more “tolerance” there is for misalignment. On the other hand, higher prior-award concentration requires greater alignment. Thus, it may be beneficial for the breadth of the prior awards to reflect the court’s confidence in their alignment with respect to the subject award.

Let us consider an example based on data obtained in Saks et al.,\textsuperscript{32} which tested the effects, with respect to variability, of providing mock jurors with certain information regarding prior awards. In one set of control conditions in which mock jurors were provided with a fact pattern (based on actual personal injury cases) involving a “high-severity injury,” a broken back, the mean and standard deviation of the award amounts determined by participants was approximately $3$ million and $4$ million, respectively.\textsuperscript{33}

Note that these values are based on amounts determined by mock jurors rather than mock juries. Also, however, “[b]ecause the distribution for the raw dollar awards was


\textsuperscript{33} See id. at 249-53.
highly variable and positively skewed, awards greater than two standard deviations above the mean were recoded to the amount at two standard deviations.” Saks et al. thereby limited the variability of the data.

Based on the data from Saks et al., we construct Figure IV.2, which assumes a correct award ($\mu_y$) of $3\text{ million}$ and judgment variability ($\sigma_y$) of $4\text{ million}$. It is intuitive to imagine an approximately “normal” distribution with almost all awards falling between 0 and $11\text{ million}$ (that is, the mean ± 2 standard deviations). We assume also that $\mu_y = \lambda_0 = 3\text{ million}$ and $\eta_0 = 0$. That is, we view the case as its own population, rather than a draw from a distribution of awards. Thus, the figure illustrates risk (defined herein as error) as a function of the mean of the prior awards (indicating alignment). We can see, for example, that if prior awards are centered at $4\text{ million}$ with standard deviation equal to $2\text{ million}$, we reduce risk by 92% by using the shrinkage estimator rather than the classical estimator; and we can expect the award to fall within the interval $74,000$ to $5.26\text{ million}$ (that is, the mean ± 2 standard deviations), rather than $0$ to $11\text{ million}$. If the prior awards are centered at $5\text{ million}$ with variability $3\text{ million}$, we reduce risk by 76.8%, relative to the classical estimator; and we can expect the award to fall within the interval $0$ to $6.85\text{ million}$. Although not shown in the figure, it can be calculated that, for prior award mean and standard deviation equal to $8\text{ million}$ and $4\text{ million}$, respectively, we reduce risk by 36%, relative to the classical estimator. Finally, if the distribution of prior awards has a mean and standard deviation equal to $3\text{ million}$ (the correct award) and $2\text{ million},

34 Id. at 249.
respectively, the variability (in terms of standard deviation) of the estimator is $0.8$ million rather than $4$ million dollars; and risk is reduced by 96%.

**Risk of Shrinkage Estimators (Black) & Risk of Classical Estimator (Gray)**

Figure IV.2: Comparison of the risk corresponding to the shrinkage estimator (black curves) and the risk corresponding to the classical estimator (gray horizontal line) plotted against the mean of the prior awards when the correct award is $3$ million (vertical black bar), and assuming $\eta_0 = 0$ (the subject case forms its own population) and judgment variability of the subject case, $\sigma_y$, is equal to $4$ million. The black curves correspond to different values of prior-award variability, which ranges from $0.2$ million to $3$ million.

Note that, although Saks et al. used mock jurors rather than juries, our choice of judgment variability—a key factor for whether prior-award information causes accuracy to increase or decrease—is likely conservative, since our choice ($4$ million) reflects the methodology in that study whereby all award amounts above two standard deviations above the mean were reduced to the amount of two standard deviations. To be safe,
however, let us illustrate a second example using half the standard deviation used above to define judgment variability. Thus, Figure IV.3 assumes a correct award ($\mu_y$) of $3$ million and judgment variability ($\sigma_y$) of $2$ million. In this example, if prior awards are centered at $4$ million with a standard deviation of $2$ million, we reduce risk by 68.75%, relative to the classical estimator. If the distribution of prior awards is centered at $1$ million with a standard deviation of $400,000$, a distribution that is concentrated around a significantly incorrect award, we nevertheless reduce risk by 7.5%, relative to the classical estimator. Finally, if the distribution of prior awards has a mean and standard deviation equal to $3$ million (the correct award) and $2$ million, respectively, we reduce risk by 75%, relative to the classical estimator. Furthermore, if the variability in this final scenario were $200,000$, the risk associated with the shrinkage estimator is even smaller—a reduction of 99.99% (corresponding to a reduction in standard deviation from $4$ million for the classical estimator to $20,000$ for the shrinkage estimator).
Figure IV.3: Comparison of the risk corresponding to the shrinkage estimator (black curves) and the risk corresponding to the classical estimator (gray horizontal line) plotted against the mean of the prior awards when the correct award is $3$ million (vertical black bar), and assuming $\eta_0 = 0$ (the subject case forms its own population) and judgment variability of the subject case, $\sigma_y$, is equal to $2$ million. The black curves correspond to different values of prior-award variability, which ranges from $0.2$ million to $3$ million.

Lastly, we construct a third figure using data from Bovbjerg et al.,\textsuperscript{35} which examined real award data by severity of injury to analyze the variability of awards for pain and suffering. The data presented here are arguably very conservative as well, since 1) the authors altogether excluded the 5\% of award values farthest from the median; 2) the data include reported incidents of additur and remittitur; and 3) the data reflect the value of

dollars in 1987.\textsuperscript{36} Thus, in Figure IV.4, we consider the example of severity level 7 (out of 9, representing severe, but not maximal, injuries), with mean and standard deviation values of approximately $2 million and $2 million. The graph again displays different curves corresponding to different levels of prior-award variability. It also illustrates the risk associated with the classical estimator as a shaded horizontal line (since the risk of the classical estimator does not depend on prior awards) at $4 million squared, which is equal to the standard deviation squared. First, we can see that if prior awards are centered at $2 million (assumed to be the correct award) with a standard deviation of $500,000, we reduce risk by 99.65%, relative to the classical estimator. If the prior awards have a mean and standard deviation equal to $500,000 and $500,000, we reduce risk by 49.83%—a milder reduction, due to the introduction of bias (but a reduction nevertheless). If the prior awards have a mean and standard deviation equal to $500,000 and $1.5 million, respectively, we reduce risk by 64%—an improvement relative to the former scenario, due to the increase in breadth, which reduces the effect of the bias. If the prior awards have a mean and standard deviation equal to $4.2 million and $200,000, respectively, we \textit{increase} risk by 18.63%, since we have a tightly bound distribution centered at a significantly incorrect award value.

On the other hand, if the prior awards have a mean and standard deviation equal to $4.2 million and $1.5 million, we reduce risk by 37.48%, since, now, the introduction of bias is lessened due to high prior-award breadth, and the beneficial effect of prior awards on award variability dominates.

\textsuperscript{36} See id. at 919-24.
Risk of Shrinkage Estimators (black) & Risk of Classical Estimator (gray)

Figure IV.4: Comparison of the risk corresponding to the shrinkage estimator (black curves) and the risk corresponding to the classical estimator (gray horizontal line) plotted against the mean of the prior awards when the correct award is $2 million (vertical black bar), and assuming $\eta_0 = 0$ (the subject case forms its own population) and judgment variability of the subject case, $\sigma_y$, is equal to $2$ million. The black curves correspond to different values of prior-award variability, which ranges from $0.2$ million to $4$ million.

Using the derivations and examples above, we state the following conclusions:

1. Prior awards that are relatively aligned with the correct award can lead to large accuracy benefits.

2. Higher breadth of prior awards leads to reduced influence of prior awards, and therefore reduced accuracy benefits—but improvements nevertheless. Considerations in determining appropriate breadth also include 1) confidence in the alignment of the prior awards, and 2) obtaining an appropriate sample size.
3. Prior awards that are misaligned—even significantly misaligned—but are of high breadth lead to improvements, but improvements that are small relative to prior awards that are aligned and are of lower breadth.

4. Increasing only the breadth of the prior awards (without affecting alignment) will not harm accuracy, but will only reduce the influence of the prior awards, and therefore their benefits with respect to accuracy.

5. Prior awards that are significantly misaligned and are of low breadth can lead to harmful effects on accuracy. However, harmful effects generally require prior awards that are tightly bound around a significantly incorrect award.

   In short, under relatively mild conditions, the shrinkage estimator outperforms the classical estimator, an adjudicated award.

5. **Conclusion**

   Claim aggregation methods may enable a court to increase the accuracy of damage awards by allowing for the sharing of information across claims. Recent papers have argued as such in the contexts of 1) sampling a proportion of claims in a class action for purposes of extrapolating awards for unsampled class claims; and 2) providing a fact-finder with prior-award information as guidance in determining awards for pain and suffering or punitive damages.

   Our goal in this paper was to examine certain implications of what can be considered a third, but closely related, form of claim aggregation called shrinkage estimation. We analyzed the accuracy benefits of shrinkage in the contexts of sampling and comparable-case guidance; and we applied it to gain a deeper understanding of the benefits
and limitations of claim aggregation generally. We began our analysis by applying shrinkage in the class action context, and by reexamining the results obtained in *Aggregating for Accuracy* to derive alternative methods, with respect to accuracy, under various legal and cost constraints. We found that shrinkage leads to greater accuracy as compared to individual adjudication, and also as compared to the sampling methods discussed in *Aggregating for Accuracy*. But it also requires relaxing certain legal constraints that are indeed applicable in many legal contexts. We therefore discussed alternatives based on various combinations of such constraints.

We then extended our analysis to the individual claim context, applying shrinkage to gain a deeper understanding of the potential benefits and limitations of comparable-case guidance. Applying certain behavioral assumptions, we derived a formula that expresses the precise conditions, in terms of the alignment and breadth of a set of prior awards, under which comparable-case guidance leads to an increase or decrease in accuracy. We then used our analysis to draw conclusions regarding the robustness of the accuracy benefits of comparable-case guidance to variations in the set of prior awards identified, and to illustrate it using a number of figures and examples.

Shrinkage is an important concept in statistics. Although it has (unsurprisingly) received little attention in law, the idea of sharing information across claims has numerous applications. It has the potential to play a significant role in the law—at least implicitly, such as with the claim aggregation methods discussed—to reduce award variability and increase reliability generally.
V. Guiding Jurors with Prior-Award Information: A Randomized Experiment*

(Coauthor: Reagan Rose)

1. Introduction

In civil jury trials, juries are frequently asked to determine awards for pain and suffering and punitive damages. But they receive scant guidance to assist them in their assessments. As a result, these damage awards are notoriously variable, undermining the law’s fairness and deterrence objectives, and causing a wide range of harms to industries and individuals.¹ Indeed, courts have long recognized and expressed concern for this “virtually unbridled discretion”² of juries, 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

A recent article, written by one of the authors, examined the benefits of comparable-case guidance (CCG), or “prior-award information”—information regarding awards in prior

---

* Hillel J. Bavli & Reagan Rose, Guiding Jurors with Prior-Award Information: A Randomized Experiment. This research benefited immensely from the guidance of Professor Donald Rubin, and from support provided by the Institute for Quantitative Social Science and the Center for American Political Studies.


comparable cases as guidance for determining damage awards—as an approach to addressing the unpredictability of awards for pain and suffering and punitive damages.\(^4\) The article concluded that CCG is generally effective in reducing unpredictability and improving the *accuracy* of awards for pain and suffering and punitive damages.\(^5\) However, the argument for CCG methods—those proposed in *The Logic of CCG* and by previous commentators—rely on certain untested behavioral assumptions.

Specifically, the primary objection to CCG methods is the following: “if our problem is the unpredictability of awards caused by a fact-finder’s inability to assess objectively the appropriate value of awards for pain and suffering or punitive damages, how is it beneficial to provide a fact-finder with information regarding damages awarded in prior cases that are separate and distinct from the present case and that presumably suffer from the same arbitrariness that we wish to address in the present case?”\(^6\) *The Logic of CCG* addressed this objection, demonstrating that, under certain assumptions regarding the behavior of a trier of fact in response to prior-award information, CCG can improve awards for pain and suffering and punitive damages under a robust range of conditions.

Thus, the objective of the current study is to examine the effect of CCG on awards for pain and suffering and punitive damages. In particular, we present and interpret the results of a factorial experiment designed to test the effects of CCG at different levels of *bias*, *variability*, and *form* of presentation, as well as interactive effects, on the *magnitude*, *spread*,

\(^4\) See generally *The Logic of CCG*.

\(^5\) See *id.* at 11-21.

\(^6\) *Id.* at 5.
and accuracy of awards for pain and suffering and punitive damages. We examine the behavior of triers of fact in response to CCG, and whether CCG improves accuracy—that is, whether its beneficial effect on the dispersion of awards outweighs any error resulting from the distortion of the magnitude of awards—under a robust set of bias, variability, and form conditions.

In summary, we find strong evidence that prior-award information improves accuracy, and that its beneficial effect on the dispersion of awards generally dominates any distortionary, or biasing, effect on awards. Furthermore, we find strong evidence that triers of fact respond to prior-award information as predicted in The Logic of CCG, and in line with the “optimal” use of such information.

Section 2 provides a brief overview of the problem—the variability of awards for pain and suffering and punitive damages—and describes a number of methods, and CCG methods in particular, proposed to address it. Section 3 explains the methodology of the present research, and the experiment in particular. Section 4 reports and interprets our results. And Section 5 concludes.

2. Addressing the Variability of Awards for Pain and Suffering and Punitive Damages with Comparable-Case Guidance

2.1 Current Methods and Proposals for Reducing Variability

Awards for pain and suffering and punitive damages are notoriously unpredictable.7 The Supreme Court, lower courts, and numerous commentators have expressed concern

---

7 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
for the variability of such awards on grounds of fairness, deterrence, and the ruinous effects of such awards on businesses and industries, and on society’s trust in the courts.\(^8\)

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

---

\(^8\) See, e.g., Exxon Shipping Co. v. Baker, 554 U.S. 471, 499 (2008) ("The real problem, it seems, is the stark unpredictability of punitive awards"); "[t]hus, a penalty should be reasonably predictable in its 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 North America, 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 7, 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 7, 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 at 5-8 (citing relevant cases and literature).
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.9 But these devices are generally inadequate as tools for addressing variability. They are applied inconsistently and with minimal standards10; they are used to address only the most extreme awards rather than variability generally; and regular use of such methods, and replacement of jury determinations with those of the court, would arguably raise significant constitutional problems and be inconsistent with fundamental principles of tort law.11 Additionally, these methods serve as a “band-aid” rather than addressing the underlying problem—that juries receive insufficient guidance in assessing awards for pain and suffering and punitive damages.12

Second, numerous jurisdictions have imposed *damage caps* to address extreme awards. Legislatures have enacted *damage caps* for certain types of awards, such as


10 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.”).

11 See Baldus, supra note 9; *The Logic of CCG* 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). Note that, although our focus is on methods that maintain the discretion of the jury, as in *The Logic of CCG*, our analysis regarding CCG methods can easily be extended to procedures in which the court has a more substantial role (on review, or in the first instance) in deciding awards for pain and suffering and punitive damages.

12 Note, *The Logic of CCG* proposes the use of CCG in addition to the procedures of additur and remittitur, and not in place of them.
punitive damages, or damage awards generally.\textsuperscript{13} However, \textit{damage caps} address only extreme cases, and only excessive awards.\textsuperscript{14} Moreover, capping awards generally 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.\textsuperscript{15} They may raise constitutional concerns as well.\textsuperscript{16}

Third, a number of commentators have proposed using awards in comparable cases as guidance for award determinations. These methods have been proposed in various forms. Some have focused on the use of comparable cases \textit{on review}, and the importance of creating a more principled approach to a trial court’s review of awards for excessiveness.\textsuperscript{17} However, although such methods improve the standards underlying a court’s review, they suffer from many of the same problems that apply to the procedures of \textit{additur} and \textit{remittitur} generally. For example, they address only extreme awards, and regular replacement of the jury’s discretion with that of the court arguably gives rise to a range of constitutional and substantive tort problems.

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

\begin{flushleft}
\footnotesize
\textsuperscript{13} See Joseph Sanders, \textit{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).

\textsuperscript{14} The Logic of CCG at 9.

\textsuperscript{15} See Sanders, supra note 13, at 509-11; The Logic of CCG at 9.


\textsuperscript{17} See, e.g., Baldus, supra note 9.
\end{flushleft}
of a case. These methods have similarly been criticized as replacing the jury’s discretion with that of the court or a legislative body altogether removed from the case.¹⁸

Some recommendations, however, have involved “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, and using the comparable-case information as guidance.¹⁹ These recommendations are based on studies demonstrating that they are effective in controlling outlying awards and variability generally (even when such information is provided as non-binding guidance).²⁰

### 2.2 Comparable-Case Guidance

*The Logic of CCG* examines a specific type of comparability analysis. It defines the term *comparable-case guidance (CCG)* as:

scheduling or comparability-analysis methods that fulfill three fundamental requirements: 1) information used as guidance must be derived from prior “comparable” cases (as opposed to, *e.g.*, damage schedules predetermined arbitrarily by a legislative body); 2) comparable-case information must be considered by the fact-finder in particular (as opposed to, *e.g.*, a reviewing court); and 3) comparable-case information must be used as guidance only (as

---

¹⁸ *The Logic of CCG* at 10; see David A. Logan, *Juries, Judges, and the Politics of Tort Reform*, 83 U. CIN. L. REV. 903, 942-43 (2015) (“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 the Restatement (Second) of Torts § 901 cmt. a (1979))).

¹⁹ *The Logic of CCG* at 3.

opposed to, e.g., imposing a range or amount that is binding on the fact-finder).\footnote{The Logic of CCG at 4.}

Numerous courts and commentators have called for such methods. Consider, for example the case \textit{Jutzi-Johnson v. United States}, which involved an appeal from an award for pain and suffering resulting from a bench trial.\footnote{263 F.3d 753 (7th Cir. 2001)} In that case, Judge Richard Posner commented on the “acute” problem of “figuring out how to value pain and suffering.”\footnote{Id. at 758.} According to Judge Posner, notwithstanding “various 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.”\footnote{Id. at 758-59.} 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.”\footnote{Id. (internal citations omitted).} 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.”

However, calls for such methods have generally failed. Indeed, there have been substantial objections to these methods on grounds that their validity relies on the ability of a court to reliably identify an appropriate set of “comparable” cases for guidance that would improve the award at hand. The fundamental issue is not whether CCG reduces variability—it is whether CCG improves awards.

The Logic of CCG thus addresses a fundamental objection: “how is it beneficial to provide a fact-finder with information regarding damages awarded in prior cases that are separate and distinct from the present case and that presumably suffer from the same arbitrariness that we wish to address in the present case?” In essence, the paper explains how “CCG methods reduce unpredictability and improve awards for pain and suffering and punitive damages by allowing for the sharing of relevant information across cases”; and

---

26 Id. (internal citations omitted). See also Roselle Wissler et al, Decisionmaking About General Damages: A Comparison of Jurors, Judges, and Lawyers, 98 Mich. L. Rev. 751 (1999) (discussing "reforms consistent with the available data," suggesting, "[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 addituit/remittitur reviews"); Chase, supra note 7, at 775, 777-90 (discussing recommendation by the ABA Action Commission to Improve the Tort Liability System to establish "tort award commissions' established to gather and report information that would be useful in 'the framing of jury instructions, the exercise of the power of addituit and remittitur, and the process of settling cases,'" id. at 775 (quoting American Bar Assoc., Report of the Action Commission to Improve the Tort Liability System 10-15 (1987)), and proposing method involving charts providing nonbinding guidance “to allow comparison with roughly similar cases in which plaintiffs' verdicts were recovered.” Id. at 777-90); Sanders, supra note 13, at 496-507 (describing proposals and studies); Logan, supra note 18, at 939-44 (discussing proposals).

27 See The Logic of CCG at 4-5.

28 Id. at 5.
how it does so with such consistency that (extraordinary circumstances aside) a court need not be concerned with whether CCG will improve accuracy in a particular case.29

The conclusions in The Logic of CCG are premised on the notion that fundamental to a good method of addressing award variability is the ability of the method to reduce, not only variability, but error: “reducing variability does not necessarily mean improving the award. For example, it is likely unwise to encourage jurors to anchor to an arbitrary value, notwithstanding associated variability benefits. In fact, variability could be zero if the court were to dictate an award without regard to the particulars of the case.”30 Thus, although “[c]ourts and scholars are correct to be concerned about the high variability associated with awards for pain and suffering and punitive damages,” “good policy requires steps toward reducing such variability only insofar as they improve awards generally.”31

The Logic of CCG, therefore, builds upon a framework for analyzing the “accuracy” of damage awards. It concludes that CCG can improve awards for pain and suffering and punitive damages under a robust set of conditions—and specifically, conditions regarding the “appropriateness” of the prior awards identified. In other words, it concludes that the validity of CCG is not very sensitive to, and therefore does not depend heavily on, the ability of a court to identify an “appropriate” set of “comparable” cases.

29 Id.
30 Id.
31 Id. at 11-12.
But its conclusions regarding the robustness of such methods rely on certain untested behavioral assumptions. The current paper, therefore, aims to test the proposition that CCG not only reduces the variability of awards, but improves awards under a robust set of conditions. If such claims are supported by the data, they may defuse the major objections to CCG methods, and may weigh heavily in favor of the use of these methods in the courtroom.

In the following subsection, we describe the framework established in The Logic of CCG for analyzing the accuracy benefits of CCG.

2.3 Reducing Variability and Error with CCG

The Logic of CCG defines the “correct” award in a case as the award that would result if the court had perfect information regarding both the law and the facts of the case.32 It explains, however, that “we neither have perfect information nor know the correct award. Instead, the court asks a jury to arrive at an award, which will serve as an estimate of the correct outcome.”33 It further characterizes the “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.).”34 Thus, “[a] single trial ... generates a sample

---

32 See id. at 12 (citing Hillel J. Bavli, Aggregating for Accuracy: A Closer Look at Sampling and Accuracy in Class Action Litigation, 14 LAW, PROBABILITY & RISK 67, 74-78 (2015) [hereinafter Aggregating for Accuracy]). Note, we can similarly define a distribution of “correct” awards reflecting, e.g., 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).

33 The Logic of CCG at 12.

34 Id. (citing Aggregating for Accuracy at 74-78 (citing 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
from the population and an *estimate* of the correct award. Call the actual award an ‘estimate,’ and the procedure that generates the estimate an ‘estimator.’ We can then define ‘error’ in terms of distance, and (equivalently) ‘accuracy’ in terms of proximity, between the estimate and the correct award.” The “reliability” of a legal procedure is then defined as the accuracy of the award that can be expected by following the procedure.

More specifically, error is defined by *mean squared error* (MSE), which can be separated into “bias” and “variance.” Bias represents systematic error, or the difference between the expected, or average, award and the “correct” award. Variance, on the other hand, is a measure of dispersion—the variability of the award around its mean. Thus, if the estimator entails a high level of variance, but is unbiased, then it will generate estimates [i.e., awards] that are highly dispersed around the correct value. In this case, we say that the estimator is “unbiased,” but that it lacks “precision.” If the estimator is “precise” but “biased,” then it generates values that are tightly centered around the wrong value—an undesirable circumstance. If an estimator is “precise” and “unbiased,” then it will generate estimates that are close in proximity to the correct value, and we say that it is “accurate.”

---

35 *Id.* (citing *Aggregating for Accuracy* at 74-78).

36 *Id.*

37 *Id.* at 13.

38 *Id.*
Therefore, error (and thus accuracy) is measured in terms of bias and variance. The Logic of CCG models awards for pain and suffering and punitive damages formally, examining two forms of variability—claim variability, reflecting substantive (factual) differences from claim to claim, and judgment variability, reflecting random variation associated with the award of a given claim. It explains that “we want the estimate to reflect the first form of variability—that meaningful factual differences (e.g., between the present case and prior cases) should result in different awards. On the other hand, we want to minimize the second form of variability—the randomness associated with the determination of damages in a given case.” It describes the benefit of CCG as follows:

Unfortunately, since we generally do not know the correct damages award, it is impossible to distinguish the former form of variability from the latter. In other words, a tradeoff arises between minimizing bias—that our estimated damages award should, on average, be as close to the correct damages award as possible, reflecting the former form of variability—and minimizing variance, the degree of randomness associated with the estimated award. Thus, in a sense, incorporating prior-award information as guidance in determining a damages award allows the fact-finder (implicitly) to strike a balance between bias and variance so as to minimize error. By accepting the possibility of introducing some bias due to material differences between the present case and prior cases, it is possible to gain far more, in terms of reliability, due to the reduction in award variability that follows from such guidance.

But what balance should the trier of fact strike between prior-award information and the facts of the case at hand? Consider two extreme scenarios described in The Logic of CCG. First, assume that the trier of fact bases its award entirely on the prior awards

---

39 Id. at 16 (citing Aggregating for Accuracy at 74-78).
40 Id. at 16-17.
41 Id. at 17.
provided to it—say, for example, that it simply applies the average of the prior awards as its award determination. In this case, there would be no error caused by random variation, since the average of the prior awards is constant, "but to the extent that the average of the prior awards is different from the correct current award ... there will be substantial error from bias." Assume, on the other hand, that the trier of fact decides that it will not even consider the prior-award information as part of its award determination—that it will determine an award based solely on the evidence. In this case, assuming the award is unbiased in the first instance, “there will be no error from bias but substantial error from variability.”

Thus, “CCG improves the reliability 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.”

Mathematically, it is possible to derive the optimal level of influence of the prior-award information. The ideal influence is based on the random variation of the award determination (i.e., judgment variability) and the variability of the prior awards (i.e., claim variability). “Higher judgment variability suggests weaker information obtained from the present case and greater influence of the prior awards; higher claim variability

\[42\] Id.
\[43\] Id.
\[44\] Id. at 18.
\[45\] Id.
suggests weaker information obtained from the prior cases (and, similarly, higher potential for bias) and greater influence of the present case."\textsuperscript{46}

For example, where judgment variability is extremely low, and multiple adjudications yield identical outcomes, "there is little to be gained from introducing prior-award information and the possibility of bias."\textsuperscript{47} Where judgment variability is high, however (as is often the case with awards for pain and suffering and punitive damages), there is much to be gained from prior-award information.\textsuperscript{48} Similarly, highly variable prior awards suggest, \textit{e.g.}, the possibility of significant differences among the set of cases—relative to each other and relative to the subject case—and less guidance for the award determination; whereas prior awards with low variability provide clearer information, and, in a sense, stronger guidance.\textsuperscript{49}

Thus, it is theoretically possible to 1) estimate claim variability, \textit{e.g.}, by computing the variance of the prior awards, 2) estimate judgment variability, \textit{e.g.}, by computing the variance of awards resulting from repeated adjudications, and 3) arrive at an award (an estimate of the "correct" award) formulaically by, \textit{e.g.}, computing a weighted average of the mean of the prior awards and the mean of the repeated adjudications, weighted by the

\textsuperscript{46} Id.

\textsuperscript{47} Id.

\textsuperscript{48} Id.

\textsuperscript{49} Id.
inverse of the variance of each, respectively. Statisticians refer to this type of estimator as a “shrinkage” estimator.

But two problems arise. First, it is difficult and costly to compute judgment variability and generally impossible to know whether a set of prior awards is “aligned” with—or has the same center as—the “correct” award in the subject case. Therefore, a court would be unable to determine the appropriate influence of the prior awards or whether CCG would be beneficial or not (or even harmful) for the award at hand in the first instance. Second, the benefits of CCG are grounded in assumptions that a trier of fact would act as predicted, and that doing so would indeed improve accuracy. But how do we know that such assumptions are justified?

The Logic of CCG addresses the first problem by explaining that it is unnecessary to compute judgment variability or whether CCG is beneficial in a particular case. That paper argues that, in light of the variability of awards for pain and suffering and punitive damages, and the robustness of the accuracy benefits of CCG under a wide range of conditions, including those involving prior awards that are not “aligned” with the “correct” award, courts should trust that, absent extraordinary circumstances, a reasonable procedure for identifying prior awards will result in CCG highly likely to improve accuracy. “Statistically (using fundamentals of ‘shrinkage estimation’),

---


incorporating prior-award information will generally improve reliability, as long as the prior awards are not tightly bound (i.e., with low variability) around a mean that is far from the correct award in the present case—in practice, as long as the prior awards are not tightly clustered around an award that reflects facts materially dissimilar to the present case.\textsuperscript{52} The paper concludes that “[i]n short, we can expect that a reasonable method for identifying prior ‘comparable’ cases will result in prior-award information that is likely to improve reliability.”\textsuperscript{53}

The second problem, regarding whether a trier of fact would act as predicted in response to CCG, and whether doing so would in fact improve accuracy in a robust range of circumstances—notwithstanding the concerns expressed above—is the subject of our examination in the current study.

Specifically, our primary objectives are to examine 1) whether prior-award information improves accuracy under a robust set of conditions related to the bias, variability, and form of prior awards, and 2) whether triers of fact respond to prior-award information as predicted, as to justify the assumptions and conclusions in The Logic of CCG.

There have been a number of studies examining the unpredictability of award determinations and methods of addressing it.\textsuperscript{54} The survey experiment by

\textsuperscript{52} The Logic of CCG at 20-21 (internal citations omitted).

\textsuperscript{53} Id. at 21.

\textsuperscript{54} See, e.g., Saks et al., supra note 20; Diamond et al., supra note 8; Bovbjerg et al, supra note 7; Baldus et al., supra note 9; Catherine M. Sharkey, Unintended Consequences of Medical Malpractice Damages Caps, 80 N.Y.U. L. REV. 391 (2005); Leebron, supra note 7. See also Sanders, supra note 13, at 496-507 (summarizing proposals and studies); Chase, supra note 7.
Professor Saks et al. is most closely related to the current study.\textsuperscript{55} That study investigated the effects of prior-award information, in various forms, on the amount awarded for pain and suffering in personal injury cases—and specifically, with respect to award variability and “distortions” in the amounts awarded.\textsuperscript{56} Prior-award forms tested in their study include an “interval,” an “average,” an “average-plus-interval,” “examples,” and a “cap” condition. The study involved three different “severity” levels, which determined the severity of the injury at issue in a fact-pattern provided to participants. Each “guidance condition” was predetermined based on the results of a pilot study. For example, the authors used the median of the pilot study to define the “average” condition, the 10\textsuperscript{th} and 90\textsuperscript{th} percentiles to define the “interval” condition, the 10\textsuperscript{th}, 50\textsuperscript{th}, and 90\textsuperscript{th} percentiles to define the “average-plus-intervals” condition, and the 10\textsuperscript{th}, 35\textsuperscript{th}, 65\textsuperscript{th}, and 90\textsuperscript{th} percentiles to define the “examples” condition.\textsuperscript{57}

In summary, the authors found that the “interval,” “average-plus-interval,” and “examples” conditions caused a reduction in variability, while distorting award amounts minimally or not at all.\textsuperscript{58} They also found that the “cap” condition performed poorly and sometimes increased the variability and size of awards.\textsuperscript{59}

\textsuperscript{55} See Saks et al., supra note 20, at 246.

\textsuperscript{56} See id. at 246-47.

\textsuperscript{57} See id. at 248.

\textsuperscript{58} See id. at 246.

\textsuperscript{59} See id. at 253.
Although the subject of the study by Saks et al., and its use of pilot data to establish guidance conditions, are similar to those of the present study, the details of the design and analysis of the present study differ substantially from those of the study by Saks et al. and other studies. Most fundamentally, our research differs from all previous studies in its focus on accuracy, as well as spread and magnitude, under a robust set of bias, variability, and form conditions. The present study is also unique in its focus on the behavioral aspects of the model presented in The Logic of CCG, including, for example, testing the effect of prior-award variability on amounts awarded.

3. Methodology

Our objective is to test the effects of prior-award information at different levels of bias, variability, and form, and their interactions, on the magnitude, spread, and, accuracy of awards for pain and suffering and punitive damages. We test these effects experimentally (i.e., by imposing an intervention) using a factorial design and the potential outcomes framework. This means that each of the participants is exposed to a “treatment combination” (or “treatment condition”)—in the form of a survey—arising from a set of “factors,” each of which can be set to a certain value, or “level,” and that we make causal conclusions based on inferences about what an outcome would be under exposure to

---


61 See generally 2^K Factorial Designs.
alternative treatment combinations. In the current section, we describe our methodology, including the details of our experimental design ("the Design") and our analysis.

### 3.1 Survey Administration

We used Amazon's Mechanical Turk to recruit and administer surveys to 5500 participants similar to individuals found on U.S. juries (U.S. citizens who are 18 or older and English speaking). Participants enrolled in the experiment sequentially and were randomly assigned to receive one of 22 surveys, each reflecting a specific treatment combination, which determined what prior-award information, if any, the participant received as guidance in responding to the survey. In each survey, participants were presented with a fact pattern, instructions, and certain guidance for determining a damages award, based on the assigned treatment combination. They were then asked to determine a

---


63 Mechanical Turk requires that all participants be 18 or older and English speaking. The U.S. citizenship requirement was difficult to enforce; but we attempted to enforced it by requiring that all participants have an IP address located within the U.S. Additionally, we provided information regarding eligibility requirements to all potential participants and asked that they participate in the survey only if they met the stated requirements. Note that we ignored some qualifications for jury service, such as the requirement that the participant not be a convicted felon.

64 To ensure roughly equal sample sizes across treatment groups, a "dynamic randomization scheme" was used. This scheme prevents any single treatment condition from having more than two more participants than any other condition at any stage of the randomization. For example, if the first two participants to enroll in the study were randomly assigned to condition 1, the next participant who enrolled in the study would be randomly assigned to one of the remaining twenty-one conditions. This scheme can be viewed as a form of "block randomization," which 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).
damages award, provide an optional explanation, and provide certain demographic information. The primary outcome in this study is the award amount.

### 3.2 Treatment Combinations

For purposes of the Design, the experiment can be understood as two separate studies: a control study where no prior award information is available and a treatment study where prior award information is provided. The control study has one experimental factor: *scenario* (2 levels). The treatment study has four experimental factors: *scenario* (2 levels), *form* (3 levels), *bias* (2 levels), and *variability* (2 levels).

*Scenario* can be set to one of two levels: *pain and suffering* or *punitive damages.*

*Form* can be set to one of three levels: *average, list, or range.* *Bias* can be set to one of two levels: *unbiased* or *biased.* And *Variability* can be set to one of two levels: *low variability* or *high variability.*

Thus, each treatment combination is characterized by the level to which each factor is set; and the treatment combination to which a participant was randomized determined the particular survey he or she received. For example, a participant randomized to the treatment combination that involves *[scenario = pain and suffering, form = list, bias = unbiased, and variability = high variability]* received a survey that asked the participant to

---

65 Participants received $.20 for their participation in the survey.

66 The optional explanation for the award amount chosen will be analyzed separately in a second study. This outcome is therefore not addressed in this study, except to confirm our expectation that survey participants provided thoughtful responses, rather than simply filling in a response without any thought.

67 This categorization is convenient for purposes of considering the experimental factors. On the other hand, as discussed further below, our results are analyzed, and most conveniently understood, within each level of *scenario.*
determine an award for pain and suffering, and provided unbiased, high variability prior-award information in the form of a list.

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

Thus, the factors define two treatment conditions in the control study and $2 \times 3 \times 2 \times 2 = 24$ treatment conditions in the treatment study, minus four inapplicable conditions that arise from “crossing” the variability factor with the average level of the form factor, for a total of $J = 2 + 24 - 4 = 22$ treatment conditions. Hereafter, we refer to the conditions within the control study as the “control conditions” and the conditions within the treatment study as the “active treatment conditions.” All conditions are explained in more detail below.

### 3.3 The Surveys

Each survey for the control conditions contains a three-paragraph stem. The first paragraph contains general instructions:

The following survey should take 5-10 minutes. Please review the information below, and provide your response and the demographic information requested. Please do not consult any outside sources. The data we collect will be used for research purposes. We will not attempt to identify you or connect your responses to your identity. If your identity becomes known to us, it will remain private.”

The second paragraph contains a short fact pattern that briefly describes a tort committed by a corporation and the consequences of the tort for its victim. The third paragraph contains the judge’s instructions for arriving at an award.

As discussed above, the control study involves only a single factor, scenario, which
has two levels, **punitive damages** and **pain and suffering**. The level of **scenario** determines the fact pattern and the judge’s instructions in the second and third paragraphs of the stem.

In particular, the second and third paragraphs of the **punitive damages scenario** state:

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.

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.”

The second and third paragraphs of the **pain and suffering scenario** state:

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.

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.

The survey then asks the participant to determine an award for pain and suffering or punitive damages. In the punitive damages scenario, the survey asks: “Please write down the dollar amount that you would award Andrew for punitive damages, as instructed by the judge.” In the pain and suffering scenario, the survey asks: “Please write down the dollar amount that you would award Emma for her pain and suffering (past and future) and her loss of capacity for enjoyment of life, as instructed by the judge.”

Finally, once a participant enters his award, the survey asks for an optional explanation for the amount awarded, as well as certain demographic information (discussed further below).68

For the active treatment conditions, the stem contains a fourth paragraph that contains certain information regarding awards in prior comparable cases. For conditions

---

68 Note, participants were not permitted to return to previous pages. Therefore, once a participant submitted his award and proceeded to the following pages (requesting an optional explanation or demographic information), he was not permitted to return to the page requesting an award amount.
involving the list form in the punitive damages scenario, the paragraph states:

Additionally, the judge informs you that in five previous similar cases juries have determined awards for punitive damages 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.

For the range form, the paragraph states:

Additionally, the judge informs you that in previous similar cases juries have determined awards for punitive damages in amounts ranging from [________ to ________]. 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.

And for the average form, the paragraph states:

Additionally, the judge informs you that in previous similar cases juries have determined awards for punitive damages in amounts averaging [________]. 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.

Analogous paragraphs are used for each level of form for the pain and suffering scenario. Further, in the actual surveys, the bracketed monetary values above are filled in with prior-award values, where the values depend on the treatment condition to which a participant is randomized—and particularly, on the levels of bias and variability associated with that treatment condition. These values are discussed in the following subsection.

3.4 Determining Prior-Award Information

Constructing the surveys to be presented to participants involved determining numerical values that would define each treatment condition. For example, constructing a survey corresponding to a treatment condition involving unbiased prior-award information in the form of an average required choosing an award value to present to participants as
the average of prior awards. We began by estimating the “correct” award in each *scenario*, defined as in *The Logic of CCG*, as the mean of the distribution of awards that would result from infinitely repeated adjudications (of the control condition) under various conditions. Once we estimated the “correct” award in each scenario, we used the estimate to define, for example, what constitutes *biased* versus *unbiased* prior-award information. We estimated the “correct” awards using the median of the control condition in each *scenario*. For a number of reasons, 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 samples in the control conditions—relative to the infinitely repeated adjudications that determine the “correct” awards. Additionally, the awards in this study are determined by *mock jurors*—not *actual juries* in a *trial governed by a judge*, all of which serve to limit the right skew of the distributions of actual awards.

Importantly, for purposes of the current study, it is not *necessary* to define and estimate a “correct” award. Our conclusions would not be weakened were we to analyze the treatment conditions 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, CCG is intended to *improve accuracy* by *reducing error* from dispersion substantially more than it adds error from distortion—whatever the initial “distortion”

---

69 See supra § 2.3.

70 See discussion regarding limitations associated with the present study, *infra* § 5.
may be. By referencing a “correct” award, we simply assume that the award determination is initially undistorted, or unbiased.

To be safe, 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 an 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, any assumption that such a bias existed would be inconsistent with the position of the Supreme Court, lower courts, and many commentators, that reducing random variation, without more (e.g., without causing bias), is beneficial. In any event, for purposes of clarity and simplicity, as well as for consistency with previous literature, we define the “correct” awards as above, and estimate them using the medians of the control group awards.

Thus, to obtain these estimates, we conducted a pilot study designed identically to the control study described above. We used Amazon’s Mechanical Turk to recruit a sample of 400 pilot participants.71 Each participant was randomly assigned to one of the two levels of scenario (pain and suffering or punitive damages), resulting in a sample of 200 participants for each scenario. As in the control conditions in the main study, pilot participants were provided no prior-award information and were asked to determine an award based on the fact pattern they received.

---

71 An additional 200 participants were surveyed for purposes of a separate analysis. The data collected for that analysis are not relevant to this paper, and therefore are not discussed.
As suggested above, participants randomized to receive one of the active treatment conditions in the main study were provided with prior-award information as a numerical summary in the form of either an average, a range, or a list of values. To determine these numerical summaries for each scenario and for each possible combination of bias and variability, we used the raw data from the pilot study. Pilot study data were then excluded from subsequent analyses.

For each level of scenario, the empirical median of award amounts in the corresponding pilot sample is used as the unbiased average and the 30th percentile of award amounts is provided as the biased average. Additionally, for each level of scenario, the 15th, 25th, 50th, 75th, and 85th percentiles of award amounts in the corresponding pilot sample are used as the unbiased high variability list, and the 40th, 45th, 50th, 55th, and 60th percentiles are used as the unbiased low variability list. Finally, for each level of scenario, the 15th and 85th percentiles of the award amounts from the corresponding pilot sample are used as the unbiased high variability range, and the 40th and 60th percentiles are used as the unbiased low variability range.

To define high variability and low variability prior-award information in the biased conditions, we apply the following procedure. First, a bias ratio is calculated as the ratio of the 30th percentile to the median within each level of scenario. For each level of scenario, we then multiply the values determined for the unbiased high variability list and unbiased low variability list by the corresponding bias ratio to construct the biased high variability list.

---

72 By the definition of “bias,” any value other than the empirical median of the corresponding pilot sample distribution could be used to define “biased” prior-award information. Here, the 30th percentile was chosen to reflect very substantial, but not entirely unrealistic, court “error” in determining a set of prior awards.
and biased low variability list, respectively. We apply the same procedure to construct the biased high variability range and biased low variability range for each level of scenario.

Table A provides the numerical prior-award values that were provided to participants in the main study in each of the active treatment conditions.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Bias and Variability</th>
<th>Form</th>
<th>Prior-Award Information$^{73}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain and Suffering</td>
<td>Unbiased Low Variability</td>
<td>Average</td>
<td>$2m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$1.12m to $4.7m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$1.12m, $2m, $2m, $3m, $4.7m</td>
</tr>
<tr>
<td></td>
<td>Unbiased High Variability</td>
<td>Average</td>
<td>$2m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$200k to $15m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$200k, $500k, $2m, $10m, $15m</td>
</tr>
<tr>
<td></td>
<td>Biased Low Variability</td>
<td>Average</td>
<td>$1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$560k to $2.35m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$560k, $1m, $1m, $1.5m, $2.35m</td>
</tr>
<tr>
<td></td>
<td>Biased High Variability</td>
<td>Average</td>
<td>$1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$100k to $7.5m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$100k, $250k, $1m, $5m, $7.5m</td>
</tr>
<tr>
<td>Punitive Damages</td>
<td>Unbiased Low Variability</td>
<td>Average</td>
<td>$1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$500k to $1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$500k, $500k, $1m, $1m, $1m</td>
</tr>
<tr>
<td></td>
<td>Unbiased High Variability</td>
<td>Average</td>
<td>$1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$10k to $10m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$10k, $52.5k, $1m, $5m, $10m</td>
</tr>
<tr>
<td></td>
<td>Biased Low Variability</td>
<td>Average</td>
<td>$100k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$50k to $100k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$50k, $50k, $100k, $100k, $100k</td>
</tr>
<tr>
<td></td>
<td>Biased High Variability</td>
<td>Average</td>
<td>$100k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range</td>
<td>$1k to $1m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>List</td>
<td>$1k, $5.25k, $100k, $500k, $1m</td>
</tr>
</tbody>
</table>

$^{73}$ We use “m” to denote millions and “k” to denote thousands.
3.5 Assessing and Correcting Covariate Balance

Due to resource and privacy constraints, covariate data were unavailable prior to the start of the experiment. In particular, data on nine covariates—age, sex, ethnicity, education, employment status, marital status, household income, residential community type, and political affiliation—were collected upon a participant’s completion of the survey.

It is desirable to have reasonable covariate balance across all treatment conditions. However, with nine covariates, there was a high probability of observing imbalance on at least one covariate among the 22 conditions. This was significant, because imbalance could prevent us from knowing whether any observed effects were attributable to the intervention or to the imbalance. Therefore, once the data were collected, and before proceeding to the general analysis phase of the experiment, we sought to examine and correct for any substantial covariate imbalance produced by the observed randomization. But, to avoid data “dredging” and ensure that covariate balance would be corrected in an objective manner, we analyzed the covariate data in a so-called “secondary design phase.” Specifically, we analyzed the data in two phases. First, we removed all outcome data and considered only covariate data. Prior to accessing the covariate data, however, we designed and finalized procedures for assessing covariate balance and for addressing any imbalances.74

To assess covariate balance across the 22 treatment groups, we examined data on the nine demographic variables measured in the survey. Additionally, we examined two

---

74 Design and Analysis Protocol on file with authors. If the observed covariate balance were not suitable, we would have proceeded by borrowing methods, such as “trimming” based on propensity scores, established for causal inference with observational data.
variables constructed using survey metadata regarding enrollment times—*time of day* and *day of week* of survey enrollment—to check for systematic differences in participants based on study enrollment times that may influence overall covariate balance and validate our assumption that participants enrolled in the study randomly.\(^75\)

The observed randomization produced treatment groups that were approximately equal in size, with the smallest treatment group containing 245 participants and the largest group containing 260 participants. Because the sample size of each treatment group was approximately equal across the 22 groups, we applied no weighting by treatment group size when assessing covariate balance.

We analyzed covariate balance using Pearson’s chi-squared test. When implementing this test in practice, it is preferable, for purposes of statistical validity, to have five or more (expected) observations in each cell of the covariate contingency table.\(^76\) If this criterion is not satisfied, it is common practice to combine, where sensible, neighboring levels of the table. We thus followed this procedure for levels of covariates that contained less than 5 observations per cell, in order to ensure that each cell would contain a sufficient number of observations. In particular, we combined certain neighboring levels

\(^75\) We also considered a third variable related to enrollment times—*weekend*—which indicates whether each participant enrolled during a weekend (defined as any time of day on Saturday or Sunday). No significant imbalances on this variable were identified, and because this indicator is a deterministic function of *day of week*, which is also shown to be balanced across all 22 treatment groups, we excluded *weekend* from further analyses.

within each of the following covariates: age, ethnicity, education, employment, income, and residential community type.\(^77\)

Independent chi-squared tests revealed no statistically “significant” differences among the 22 treatment groups in the frequency distribution of each of the nine demographic variables (including age \(p = 0.15\), sex \(p = 0.55\), ethnicity \(p = 0.82\), education \(p = 0.33\), employment \(p = 0.87\), marital status \(p = 0.76\), income \(p = 0.34\), residential community type \(p = 0.34\), and political affiliation \(p = 0.71\)).\(^78\) Similarly, no significant differences among treatment groups were identified for either of the enrollment-time variables (including day of week \(p \approx 1\) and time of day \(p \approx 1\)). The observed \(p\)-value for the chi-squared test on each of the enrollment-time variables supports our assumption that participants enrolled in the study randomly and provides evidence that the randomization scheme performed as intended.

\(^77\) For age, we aggregated levels for ages above 61 to create one “Over 61” level. For ethnicity, participants who marked “American Indian or Alaska Native” were aggregated with participants marking the value, “Other.” For education, participants who marked “None,” “Some high school,” or “High school graduate” were aggregated into the level of “High school or less,” and participants who marked “Master’s degree,” “Doctorate degree,” “Law degree,” or “Medical degree” were aggregated into the level, “Master’s degree or higher.” For employment status, participants who marked “Retired” were aggregated into the level, “Other.” For income, levels “$100,000 - $150,000” and “Over $150,000” were aggregated into the level, “Over $100,000.” Finally, for residential community type, participants who marked “Suburban” or “Don’t know” were aggregated into the level “Suburban or uncertain.” Additionally, for the covariate sex, we observed 20 participants who marked the response “Other” (averaging one or fewer observations per treatment group). Because there is no clear way to recode these observations into either the “Male” or “Female” levels, we excluded these participants from subsequent analyses. If there are systematic differences in response patterns for the subpopulation that identifies as “Other,” the exclusion of these observations may limit our ability to generalize results from this study to this subpopulation. However, we performed post-hoc analyses to evaluate the impact of excluding these participants from the study and to understand the generalizability of our results. One participant who identified as “Other” was removed in the data cleaning process (described below); and correlation analysis of the remaining 19 observations indicates no significant associations between identification as “Other” and award amount, both within and across levels of scenario. This suggests that removal of these 19 participants from our analysis has no considerable implications on the generalizability of the study.

\(^78\) We present here raw, unadjusted \(p\)-values.
To further confirm suitable balance for relevant comparison groups, we performed pairwise comparisons to evaluate differences in frequency distributions on each covariate for each of the \( \binom{22}{2} = 231 \) pairs of distinct treatment groups using Fisher’s exact test.\(^7\) To correct for multiple comparisons, we applied the procedure proposed by Benjamini and Hochberg (1995), herein referred to as the “Benjamini-Hochberg procedure,” to control the false discovery rate (FDR).\(^8\) After adjustment for multiple comparisons, no significant differences between pairs of treatment groups were identified on any of the eleven variables tested.

We also performed tests for two-way interactions among the nine demographic variables and the two enrollment-time variables. These interaction effects can help us to better understand the observed sample and evaluate how inference from this study generalizes to the population of interest. Again, we employed the Benjamini-Hochberg procedure to adjust for multiple comparisons. After adjustment, of the \( \binom{11}{2} = 55 \) tests performed, 46 interactions were identified as statistically “significant” at the overall level of \( \alpha=0.05.\(^8\) Such correlations are not surprising (for example, age may correlate with experience and/or responsibility, and thus with employment; day of week may correlate

\(^7\) We also performed comparisons across aggregations of treatment groups corresponding to the 60 effects of interest defined in our primary analysis and found suitable balance for each aggregation.


\(^8\) For example, sex significantly correlates with marital status, age significantly correlates with employment status, and day of week significantly correlates with time of day.
with lighter work schedules and thus with *time of day* of survey enrollment). Of greater interest are the interactions of the demographic variables with the enrollment-time variables, since extreme correlations of this type may have implications regarding the assignment mechanism.\footnote{For example, if females enrolled in the study exclusively on Mondays, the number of randomizations that are possible to observe using our randomization scheme would decrease, having implications for our analysis for purposes of making valid inferences.} In these tests, significant interactions were identified for *time of day* and the demographic variables *sex*, *income*, and *marital status*. However, because we observed suitable balance on each of the eleven variables across each of the 22 treatment groups, there was no reason to believe that the identified interactions may have introduced confounding into the randomization or have negative implications with respect to the validity of the study. These interaction effects simply serve to characterize the population represented by our sample and may be further considered in post-hoc analyses and/or when attempting to generalize the results of the study.

In summary, our analysis of the covariate data indicated that suitable balance on all demographic variables and enrollment-time variables across the 22 treatment groups was achieved by the observed randomization. Therefore, we proceeded to the analysis phase of the experiment without applying any corrections or adjustments to improve balance.

### 3.6 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. We therefore base our analyses and estimation of causal effects on the potential outcomes framework of Neyman, often referred to as the Rubin Causal Model (RCM). We follow 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.

Under the RCM, each unit, i.e., each participant, in this experiment has 22 “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, $1 million had he been randomized to the biased, low variability, average, punitive damages treatment condition, and so on and so forth for all 22 possible treatment combinations. The RCM frames causal inference as a missing-data problem: because we can observe only 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.

---

83 For example, one drawback of the linear model framework, as suggested in $2^K$ Factorial Designs, 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.


86 See generally id.; Rubin, supra note 60; $2^K$ Factorial Designs.
In general, a factorial “main effect” of an experimental factor for each unit is the difference in a function (e.g., the mean) of the potential outcomes between a subset of the potential outcomes corresponding to one level of the factor and the remaining (or, herein, a subset of) potential outcomes. For example, the unit-level main effect of unbiased prior-award information relative to biased prior award information is the difference between a function (such as the mean) of the potential outcomes associated with the subset of treatment combinations that include unbiased prior-award information (that is, the outcomes that would occur had the unit been assigned to these treatment combinations) and the same function of the potential outcomes associated with the subset of treatment combinations that include biased prior-award information (that is, the outcomes that would occur had the unit been assigned to these treatment combinations).

Notationally, as in $2^K$ Factorial Designs, let $K$ denote the number of experimental factors, which are indexed by $k$, each with $p_k$ levels, such that there are $J = p_1 \times \cdots \times p_K$ treatment combinations. Let $z_j$ for $j=1,...,J$ denote one particular treatment combination, defined by a $K$-dimensional vector, where the $k$th element indicates the level of the $k$th factor, for $k=1,...,K$. Let $Z$ denote the set of all possible treatment combinations. Further, let $N$ denote the total sample size, indexed by $i=1,...,N$. And let $Y_{i}(z)$ denote the potential outcome of the $i$th unit under exposure to treatment $z$. $2^K$ Factorial Designs at 730. Let $Y_i=(Y_{i1},...,Y_{iJ})$ be the vector of all potential outcomes for unit $i$. Define $g$ as the $J$-1 vectors, indexed by $j=1,...,J-1$, such that $g_j$ is a vector of dimension $J$ with values that indicate the levels of factors corresponding to potential outcomes. Id. For example, if we refer to scenario as factor 1, then $g_j$ is a vector of eleven -1’s followed by eleven +1’s, indicating that the first eleven treatment combinations $z_{1},...,z_{11}$ have the scenario factor set to pain and suffering and the following eleven combinations have the scenario factor set to punitive damages. Using this notation, the factorial main effect of treatment combination $j$ compared to all other treatment combinations for unit $i$ can be defined generally as

$$\tau_{ij} = \frac{1}{\#\{g_j = 1\}} \sum_{g_j = 1} Y_{i} - \frac{1}{\#\{g_j \neq 1\}} \sum_{g_j \neq 1} Y_{i} \text{ where } \#\{g_j = 1\} \text{ denotes the number of elements of } g_j \text{ that are equal to one.}\$$

Similarly, the factorial effect of treatment combination $j$ relative to treatment combination $j'$ for unit $i$ can be defined as

$$\tau_{ij,j'} = \frac{1}{\#\{g_j = 1\}} \sum_{g_j = 1} Y_{i} - \frac{1}{\#\{g_{j'} = 1\}} \sum_{g_{j'} = 1} Y_{i} \text{. Factorial effects for each level } p_k \text{ of each factor } k \text{ are defined by aggregating over treatment combinations that have factor } k \text{ set to that level. These definitions are similarly extended to subsets of treatment combinations. See id.}$$
Two-way, three-way, and four-way “interaction effects” can be defined similarly. For example, the unit-level effect of unbiased high variability prior-award information, relative to unbiased low variability prior-award information, can be defined as the difference between the mean of the potential outcomes associated with the subset of treatment combinations that include unbiased high variability prior-award information and the mean of the potential outcomes associated with the subset of treatment combinations that include unbiased low variability prior-award information.88

Note that we do not test “across” levels of scenario. That is, for purposes of analysis, we treat the pain and suffering scenario and the punitive damages scenario as separate studies; and all comparisons, corresponding to effects of interest, are made within levels of scenario—that is, within the pain and suffering scenario or the punitive damages scenario. Thus, we sometimes refer to, e.g., the effect of bias or variability, within a level of scenario, as a “main effect.”

3.7 Estimands and Estimators

The primary objectives of the experiment are to assess the finite-population effects of prior-award information, for different levels of bias, variability, and form, on the magnitude, spread, and accuracy of resulting award values, where accuracy is defined as the

---

88 Throughout the study, we make 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. 2K Factorial Designs at 730. For this experiment, we have no reason to believe that there was interference across units because the population from which we recruited participants consists of many distinct users from across the United States. Further, note that Mechanical Turk required that each survey participant have a unique IP address. Similarly, because we carefully controlled the administration of all surveys, we are arguably justified in assuming that there were no hidden versions of treatment. We assume that each participant read the survey they were assigned in its entirety.
proximity of the awards to the “correct” award in terms of both bias and variance. These objectives motivate our choice of “estimands,” the quantities we are interested in estimating. We define three general estimands—one for magnitude, one for spread, and one for accuracy—and corresponding “estimators,” functions of the data that we use to estimate the quantities of interest, i.e., the estimands.

Magnitude can be captured quantitatively using a measure of central tendency, such as the mean or median. Because the outcome is monetary and known to be positively skewed, for purposes of analysis we define magnitude as the mean of the logarithm-transformed award amounts. Although this definition involves a number of drawbacks, discussed below, it reduces the effect of the right skew and it allows us to capture relative, rather than absolute, differences. On the other hand, we define spread as the interquartile range (IQR) of the outcome, which is more robust to outliers than the variance; and we define accuracy as the inverse of the mean-squared error (MSE), which is equal to the sum of the variance and the squared bias. Thus, to calculate accuracy, we establish a numerical rule for defining the “correct” award, reflecting our estimator for the “correct” award discussed supra § 3.4. Specifically, within each level of scenario, the median amount awarded in the control group of the main study is used as the “correct” award.

We thus define the finite-population factorial effect of each experimental factor and interactions between factors on the magnitude of award values as the difference between

---

89 For example, the difference between $0 and $100 should arguably be interpreted differently than the difference between $1,000,000 and $1,000,100. Although capturing relative differences is important, this definition may not be ideal or appropriate in the present context. We discuss this, and alternative definitions, infra § 4.4.
the mean of the logarithm-transformed potential outcomes associated with each of the comparison groups, respectively. We apply this definition to all such comparison groups, whether for main effects, or two-way, three-way, or four-way interaction effects. For example, in the pain and suffering scenario, the effect on magnitude of low variability unbiased prior-award information, relative to high variability unbiased prior-award information, is the difference between the mean of the logarithm-transformed potential outcomes associated with the former and the mean of the logarithm-transformed potential outcomes associated with the latter.

We estimate these effects (since the estimand is a function of unknown quantities) by using the corresponding sets of observed data. We define the finite-population factorial effect on the spread of award values as the interquartile range (IQR)—defined as the difference between the 75th percentile and the 25th percentile—of the potential outcomes associated with each of the comparison groups, respectively. We estimate this effect using the corresponding sets of observed data.

---

90 More formally, we define the finite-population factorial effect of treatment combination $j$ compared to all other treatment combinations on the magnitude of award values as $\bar{Y}^{\text{max}} = \frac{1}{\#\{g_j = 1\}} \sum_{g_j = 1} Y - \frac{1}{\#\{g_j \neq 1\}} \sum_{g_j \neq 1} Y$, where $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \log(Y_i + 1)$ denotes the mean of the logarithm-transformed potential outcomes. We estimate this quantity using the estimator $\hat{Y}^{\text{obs}} = \frac{1}{\#\{g_j = 1\}} \sum_{g_j = 1} Y^{\text{obs}} - \frac{1}{\#\{g_j \neq 1\}} \sum_{g_j \neq 1} Y^{\text{obs}}$. As above, these definitions, as well as the definitions for spread and accuracy below, are similarly extended to comparisons of particular treatment combinations or subsets of treatment combinations to other particular treatment combinations or subsets of treatment combinations. See supra note 87. See generally 2^K Factorial Designs.

91 We define the finite-population factorial effect of treatment combination $j$ compared to all other treatment combinations on the spread of award values as
Finally, we define the finite population factorial effect on the error of award amounts, the inverse of the accuracy of award amounts—and thus on the accuracy of award amounts itself—as the difference between the MSEs, defined as the mean of the squared differences between each value and the “correct” value, of the potential outcomes associated with each comparison group, respectively. Here, again, we estimate the effect on accuracy using the observed data corresponding to the comparison groups.92

Using the factorial-effect estimators we estimate an effect by computing the values of magnitude, spread, or accuracy for each of the treatment combinations involving the factor (or combination of factors) of interest, and then separately averaging over each of

\[
\tau_{\text{spread}} = \frac{1}{\#(g_j = 1)} \sum_{g_j=1} (q^{0.75}(Y) - q^{0.25}(Y)) - \frac{1}{\#(g_j \neq 1)} \sum_{g_j \neq 1} (q^{0.75}(Y) - q^{0.25}(Y)), \text{ where } q^i(Y) \text{ denotes the } i\text{th quantile of the empirical distribution of } Y. \text{ This effect is estimated using the estimator }
\]

\[
\hat{\tau}_{\text{spread}} = \frac{1}{\#(g_j = 1)} \sum_{g_j=1} (q^{0.75}(Y_{\text{obs}}) - q^{0.25}(Y_{\text{obs}})) - \frac{1}{\#(g_j \neq 1)} \sum_{g_j \neq 1} (q^{0.75}(Y_{\text{obs}}) - q^{0.25}(Y_{\text{obs}})). \text{ See generally 2K Factorial Designs.}
\]

92 We define the finite-population factorial effect of treatment combination \( j \) compared to all other treatment combinations on the accuracy of award amounts as

\[
\tau_{\text{acc}} = \frac{1}{\#(g_j = 1)} \sum_{g_j=1} (Y - \alpha)^2 - \frac{1}{\#(g_j \neq 1)} \sum_{g_j \neq 1} (Y - \alpha)^2,
\]

where \( \alpha \) is the “correct” award, which takes on a different value for each scenario. If \( z_1 \) represents the control condition in the pain and suffering scenario and \( z_{12} \) represents the control condition in the punitive damages scenario, we have \( \hat{\alpha}^{PS} = \text{med}(Y_{\text{obs}}(z_1)), \) as the estimator of the “correct” award for the pain and suffering scenario and \( \hat{\alpha}^{PD} = \text{med}(Y_{\text{obs}}(z_{12})) \) as the estimator of the “correct” award for the punitive damages scenario. We then estimate the factorial effect of treatment combination \( j \) on the accuracy of award amounts for the pain and suffering level of scenario using the estimator

\[
\tau_{\text{acc}, PS} = \frac{1}{\#(g_j = 1)} \sum_{g_j=1} (Y_{\text{obs}} - \hat{\alpha}^{PS})^2 - \frac{1}{\#(g_j \neq 1)} \sum_{g_j \neq 1} (Y_{\text{obs}} - \hat{\alpha}^{PS})^2, \text{ and for the punitive damages level of scenario using the estimator }
\]

\[
\tau_{\text{acc}, PD} = \frac{1}{\#(g_j = 1)} \sum_{g_j=1} (Y_{\text{obs}} - \hat{\alpha}^{PD})^2 - \frac{1}{\#(g_j \neq 1)} \sum_{g_j \neq 1} (Y_{\text{obs}} - \hat{\alpha}^{PD})^2, \text{ where } g_j^{PS} \text{ and } g_j^{PD} \text{ are vectors with values indicating the level of each factor for each potential outcome when the scenario factor is set to pain and suffering or punitive damages, respectively. See generally id.}
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) calculate the magnitude of award amounts separately for each of the 10 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) average the values from step (1) across the 5 treatment combinations receiving biased information, and separately average the values from step (1) across the 5 treatment combinations receiving unbiased information; and 3) calculate the difference of the averages obtained in step (2).

Also note that the estimands and estimators defined in this subsection apply to main effects, as well as all two-way, three-way, and four-way interaction effects.

### 3.8 Analysis Procedure

Once we calculate the estimates of the factorial effects using observed data, we are interested in understanding 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. The RCM allows such inference using the Neymanian (1923, 1990) perspective, Fisherian perspective (1925, 1935), or the Bayesian perspective.

Because the estimator for capturing, e.g., the spread of award values is not an unbiased estimator of its corresponding estimand, and for other reasons, we decided not to

---

93 Rubin, supra note 60.

use Neyman’s method. Instead, we adopt the Fisherian approach, and use randomization tests to evaluate Fisher’s sharp null hypothesis of “no unit-level treatment effects.” For example, to test whether the estimate of the main effect of each factor on magnitude, spread, and accuracy of award values is “significant,” we assume Fisher’s “sharp null”—a hypothesized effect, such as a “sharp null” of “zero treatment effect”—in order to impute the missing potential outcomes for each unit and repeatedly calculate the value of a specified test statistic under different possible treatment assignments. Then, given the distribution of each test statistic, we calculate approximate Fisher “exact p-values” (FEPs) to indicate the statistical “significance” of effects. We can also use this procedure to construct Fisher exact intervals for each estimate.

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.

Due to limited computational resources, we follow the common approach of constructing randomization distributions using samples of the possible randomizations, rather than computing all possible randomizations of the N=5480 participants into the 22 treatment conditions.

Due to the number of experimental factors and questions of interest in this study, to estimate and evaluate the statistical “significance” of each of the effects of interest requires numerous hypothesis tests. When testing many hypotheses, the probability of identifying false effects increases with the number of tests performed. To adjust for this increased
probability of false identifications, it is appropriate to apply an adjustment, or “correction,” to all estimated $p$-values. However, application of such an adjustment across a large number of tests generally decreases the power of the study to identify results that are actually “significant.”

To address this issue, we divide our tests into a primary analysis category and a secondary analysis category, corresponding to the effects of primary and secondary interest, respectively, and apply corrections for multiple comparisons separately within each category. The tests in the primary category relate to the study’s primary questions of interest. The tests in the secondary category relate to 1) questions that are of secondary interest, 2) effects of interest that may provide confirmatory or supportive evidence of results identified in the primary analysis category, or that may answer questions that follow from results identified in the primary analysis category, or 3) exploratory analysis.

Because we are most interested in understanding primary effects within each level of scenario, in the primary analysis category we treat as individual experiments 1) all treatment conditions receiving the pain and suffering level of scenario, and 2) all treatment conditions receiving the punitive damages level of scenario. Thus, in the primary analysis category, we test each effect of interest within each of the two levels of scenario and apply a correction for multiple comparisons separately within each. In particular, we test ten effects within each of the two levels of scenario on each of the three outcomes of interest ($magnitude$, $spread$, and $accuracy$), for a total of $10 \times 2 \times 3 = 60$ primary hypothesis tests.\footnote{In the primary analysis category, we test the following ten tests within each level of scenario: 1) the effect of unbiased prior-award information relative to no prior-award information (control); 2) the effect of $unbiased$ prior-award information relative to no prior-award information (control); 3) the effect of $biased$ prior-award information relative to no prior-award information (control); 4) the effect of $unbiased$ prior-award information relative to $biased$ prior-award information (control); 5) the effect of $unbiased$ prior-award information relative to no prior-award information (control); 6) the effect of $unbiased$ prior-award information relative to $biased$ prior-award information (control); 7) the effect of $unbiased$ prior-award information relative to no prior-award information (control); 8) the effect of $unbiased$ prior-award information relative to $biased$ prior-award information (control); 9) the effect of $unbiased$ prior-award information relative to no prior-award information (control); 10) the effect of $unbiased$ prior-award information relative to $biased$ prior-award information (control).}

\footnote{See Benjamini & Hochberg, supra note 80.}
All additional hypothesis tests are performed in the secondary analysis category, including tests for the main effects of form and variability, and additional two-way, three-way, and four-way interaction effects among bias, variability, and form. In both analysis categories, we correct for multiple comparisons as we did in the Secondary Design Phase in the analysis phase of this study, using the Benjamini-Hochberg procedure to control the false-discovery rate (FDR).\(^9^7\)

Our procedure for estimating factorial effects and carrying out inference for each, which includes computing \(p\)-values to test the null hypotheses of “no treatment effect” and constructing intervals for the estimated effects, is outlined below.

**Estimating effects.** For each estimand of interest, we use the observed data to compute the value of the corresponding estimator. Because the estimands are defined as differences between treatment groups, for each estimate, we are interested in testing the null hypothesis of “no difference between groups”—that is, we are interested in testing the

\[ unbiased \text{ prior-award information relative to no prior-award information (control) when prior-award information is low variability; } 3) \text{ the effect of unbiased prior-award information relative to no prior-award information (control) when prior-award information is high variability; } 4) \text{ the effect of biased prior-award information relative to no prior-award information (control); } 5) \text{ the effect of biased prior-award information relative to no prior-award information (control) when prior-award information is low variability; } 6) \text{ the effect of biased prior-award information relative to no prior-award information (control) when prior-award information is high variability; } 7) \text{ the effect of biased prior-award information relative to unbiased prior-award information when prior-award information is high variability; } 8) \text{ the effect of biased prior-award information relative to unbiased prior-award information when prior-award information is low variability; } 9) \text{ the effect of low variability prior-award information relative to high variability prior-award information when prior-award information is unbiased; } 10) \text{ the effect of low variability prior-award information relative to high variability prior-award information when prior-award information is biased.} \]

\(^9^7\) This procedure limits the expected proportion of tests falsely identified as “significant” relative to the total number of tests identified as “significant.” It is a well-accepted approach to correcting for multiple comparisons while often resulting in more power than methods controlling the familywise error rate (FWER), such as the Bonferroni correction and the ensemble adjusted \(p\)-value approach. See Maria T. Kimel et al., *The False Discovery Rate for Multiple Testing in Factorial Experiments*, 50(1) Technometrics 32 (2012).
hypothesis that the true difference between groups is zero. Therefore, for each hypothesis test, we use the corresponding estimator to define the test statistic for our randomization test.

**Constructing randomization distributions.** For each effect of interest, we assume the sharp null hypothesis of “no unit-level treatment effect” to impute missing potential outcomes for each participant. We then generate a sample of \( N_{\text{sim}} = 250,000 \) possible randomizations of the \( N \) participants.\(^98\) For each of these 250,000 possible randomizations, we recompute the value of the test statistic under the (hypothetical) randomization. The resulting distribution is referred to as the “randomization distribution” of the test statistic.

**Testing hypotheses.** We approximate FEPs for each hypothesis by calculating the proportion of values in the randomization distribution that were equal to or more extreme than the observed test statistic. All estimated \( p \)-values are adjusted for multiple comparisons using the Benjamini-Hochberg procedure to control the FDR. After correction, we consider estimated \( p \)-values of less than \( \alpha = 0.05 \) to be “statistically significant.” Significant \( p \)-values suggest that there is substantial evidence within the data against the null hypothesis of no treatment effect.

**Constructing intervals.** For each factorial effect, we can generate a 95% Fisher interval by calculating a sequence of raw, unadjusted \( p \)-values corresponding to a null hypothesis other than the hypothesis of “zero treatment effect,” and then identify the

---

\(^98\) We choose \( N_{\text{sim}} = 250,000 \) to ensure precision of the estimated \( p \)-value. Professors Imbens and Rubin show that, for a sample of this size, the standard error of each estimated \( p \)-value will be less than 0.001. See Guido W. Imbens & Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences* (2015).
hypothesized values that result in $p$-values equal to or larger than 0.05. For all estimates we assume a constant additive treatment effect.\textsuperscript{99}

4. Results

Before discussing our results, it is important to note that, prior to accessing any data, we drafted and finalized an *Experimental Design and Analysis Protocol* that detailed the design of the study and all aspects of our intended analysis.\textsuperscript{100} We ensured that we would not have access to the data by asking a third-party administrator to lock our access on the relevant database, and to unlock the data, when instructed, only upon our circulation (and posting to SSRN) of a finalized *Experimental Design and Analysis Protocol*.

In this section, we report and interpret our results. For purposes of clarity and ease of interpretation, we divide the section into topics, reflecting categories of effects of interest. For each topic, we report our results and accompanying interpretation. Then, in the following section, the Conclusion, we tie our findings together, relate them to the model introduced in *The Logic of CCG*, and discuss limitations.

As a preliminary matter, it is important to realize that our analysis involved approximately 250 hypothesis tests, reflecting main effects and interactions among the various factors. There can be numerous tests that speak to a particular effect, and some results may seem unsupportive of, or even inconsistent with, others. Our general approach to interpreting the data is as follows: First, we examine the comparison that most directly

\textsuperscript{99} Note, because we use the logarithm transformation for estimating factorial effects on magnitude, we interpret the resulting intervals for these estimates as multiplicative effects.

\textsuperscript{100} On file with the authors.
speaks to an effect of interest. Second, we examine other comparisons that speak to the effect less directly. Third, if we find support for the effect in step (1) and support for the effect in step (2), we generally interpret the data as evidencing a strong effect. If we find support for the effect in step (1) but an absence of evidence for or against the effect in step (2), we generally interpret the data as evidencing an effect, unless there is a particular reason to believe otherwise. If we find support for the effect in step (1) but evidence against the effect in step (2), we examine the inconsistency that arose in step (2). If our examination reveals a straightforward explanation that defuses the inconsistency, we interpret the data as evidencing an effect. If the examination fails to reveal a straightforward explanation, we interpret the data as not evidencing an effect; and we examine the results for an alternative explanation.

4.1 The Data

The data can be divided into two datasets: one for the punitive damages scenario and one for the pain and suffering scenario.\footnote{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 23 participants who entered an incorrect payment code that was intended to indicate, for purposes of receiving payment, that the participant completed the survey. 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 there was no evidence suggesting any resulting systematic distortion, we excluded these participants without applying advanced missing data techniques. Additionally, as discussed in the Methodology Section, 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 enforcing these exclusion criteria, our final sample size is N=5,458.} There are approximately 240-250 observations in each treatment level for each scenario. The award determinations are extremely variable. Awards range from $0 to $15 billion in the punitive damages scenario, with a 99th percentile
of $50 million, and $0 to $1 Quadrillion ($1 \times 10^{15}$) in the *pain and suffering scenario*, with a 99\textsuperscript{th} percentile of $100 million.

To address extreme outliers, we “truncate” the data in the *pain and suffering dataset* (in which the problem is particularly acute) at the 99\textsuperscript{th} percentile. This means that any awards above $100 million are recoded to $100 million. This is important, since extreme outliers can cause spurious results that are dominated by chance rather than true effects.\textsuperscript{102}

We applied this approach to the *punitive damages* data also, but find that the truncation method in that *scenario* has no effect on our findings, because the outliers in the *punitive damages scenario* are far less extreme in size and effect.

Thus, unless stated otherwise, the results we report and interpret below reflect the raw dataset for the *punitive damages scenario* and the *truncated* dataset for the *pain and suffering scenario*. We sometimes use the truncated data for the *punitive damages scenario* or the raw data for the *pain and suffering scenario*, but only for illustrative purposes—i.e., to make a point regarding those data in particular or, in certain instances, for purposes of creating easily-interpretable displays. In such instances, we state clearly that we are using these alternative datasets. Below, we also examine the effects of prior-award information

\textsuperscript{102} Note that, in designing the experiment, we were sensitive to the fact that, while many studies use liberal methods for addressing outliers, extreme values are of particular interest in the current study. As a result, we attempted to avoid the use of such techniques. However, because the study involves monetary values that participants chose from their imagination, the outliers in this study—and particularly those in the *pain and suffering scenario*—were particularly extreme, exceeding even our expectations. Upon observing the magnitude of the awards, we understood that not addressing the outliers may result in spurious effects and other anomalies that would result from randomness rather than a true effect. We therefore decided to address outliers in the *pain and suffering scenario* using the conservative method described in the text. We proceed using the truncated dataset. Note that such extreme awards (and likely some that are far less extreme) are highly unlikely in practice, and would, in any event, be addressed by the courts using the device of remittitur.
on outliers separately, as well as the robustness of our results using an alternative truncation threshold. Summary statistics for each scenario (raw and truncated) are provided in Table B.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Punitive Damages (Raw)</th>
<th>Punitive Damages (Truncated at 99th percentile ($50 million))</th>
<th>Pain and Suffering (Raw)</th>
<th>Pain and Suffering (Truncated at 99th percentile ($100 million))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>2751</td>
<td>2751</td>
<td>2704</td>
<td>2707</td>
</tr>
<tr>
<td>Mean (millions)</td>
<td>$18.3</td>
<td>$2.3</td>
<td>$23.8</td>
<td>$6.7</td>
</tr>
<tr>
<td>Median (millions)</td>
<td>$0.5</td>
<td>$0.5</td>
<td>$3.0</td>
<td>$3.0</td>
</tr>
<tr>
<td>Standard Deviation (millions)</td>
<td>$397.3</td>
<td>$6.4</td>
<td>$396.0</td>
<td>$14.0</td>
</tr>
<tr>
<td>IQR (millions)</td>
<td>$0.9</td>
<td>$0.9</td>
<td>$3.5</td>
<td>$3.5</td>
</tr>
</tbody>
</table>

### 4.2 Effect of Prior-Award Information 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 has a significant positive effect on *accuracy*. Figure V.1 illustrates the approximate randomization distributions and observed test statistics for the hypothesis tests of “no treatment effect” in each scenario; and effects on *error* (the inverse of *accuracy*) within each level of scenario are summarized in Table C and

---

103 See infra § 4.6.

104 We exclude from this summary of raw *pain and suffering* data three extreme values that are greater than or equal to $100 billion. These observations are removed from this summary solely for descriptive purposes.
Figure V.2 below. The dataset for the *punitive damages scenario*, in addition to the dataset for the *pain and suffering scenario*, is truncated in this subsection for purposes of interpretability and illustration (although our findings would not change were we to use the data without truncation). In Table C, Fisher intervals are provided for the difference in error between each treatment level and control, where the MSE in the control group for the *punitive damages scenario* is $146$ trillion and the MSE in the control group for the *pain and suffering scenario* is $492$ trillion.

![Punitive Damages](image1.png) ![Pain and Suffering](image2.png)

Figure V.1. Randomization distribution for effect of treatment (any prior-award information) versus control (no prior-award information) on accuracy for *punitive damages* data (left) and *pain and suffering* data (right), each truncated at its 99th percentile ($50$ million and $100$ million, respectively). Red lines show observed test statistics.
<table>
<thead>
<tr>
<th>Effect</th>
<th>MSE (trillion of dollars)</th>
<th>Difference from control (MSE (= 146)) (trillions of dollars) (95% Fisher Interval)</th>
<th>Unadjusted (p)-value(^{105})</th>
<th>MSE (trillion of dollars)</th>
<th>Difference from control (MSE=492) (trillions of dollars) (95% Fisher Interval)</th>
<th>Unadjusted (p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (any prior-award information)</td>
<td>33</td>
<td>-114 (-141, -80)</td>
<td>0.000***</td>
<td>192</td>
<td>-300 (-423, -156)</td>
<td>0.008**</td>
</tr>
<tr>
<td>Unbiased prior-award information</td>
<td>56</td>
<td>-103 (-132, -70)</td>
<td>0.000***</td>
<td>221</td>
<td>-271 (-404, -122)</td>
<td>0.002***</td>
</tr>
<tr>
<td>Unbiased low variability prior-award information</td>
<td>19</td>
<td>-127 (-161, -91)</td>
<td>0.000***</td>
<td>230</td>
<td>-262 (-415, -101)</td>
<td>0.008**</td>
</tr>
<tr>
<td>Unbiased high variability prior-award information</td>
<td>57</td>
<td>-89 (-122, -52)</td>
<td>0.001**</td>
<td>261</td>
<td>-231 (-381, -71)</td>
<td>0.013*</td>
</tr>
<tr>
<td>Biased prior-award information</td>
<td>22</td>
<td>-124 (-153, -91)</td>
<td>0.000***</td>
<td>163</td>
<td>-329 (-460, -183)</td>
<td>0.000***</td>
</tr>
<tr>
<td>Biased low variability prior-award information</td>
<td>24</td>
<td>-123 (-157, -86)</td>
<td>0.000***</td>
<td>131</td>
<td>-361 (-515, -195)</td>
<td>0.000***</td>
</tr>
<tr>
<td>Biased high variability prior-award information</td>
<td>30</td>
<td>-116 (-151, -79)</td>
<td>0.000***</td>
<td>109</td>
<td>-383 (-532, -221)</td>
<td>0.001***</td>
</tr>
</tbody>
</table>

\(^{105}\) Throughout this section, stars indicate statistical significance after correction for multiple comparisons. In particular, *** denotes statistical significance at the \(\alpha=0.001\) level, ** denotes significance at the \(\alpha=0.01\) level, and * denotes significance at the \(\alpha=0.05\) level.
Figure V.2. Observed mean squared error (MSE) (in trillions) for different treatment combinations for punitive damages (left) and pain and suffering (right). Dotted lines show MSE for control group, and stars indicate statistical significance (after correction for multiple comparisons) of the difference in MSE between each treatment combination and control.

As discussed above, prior-award information is expected to cause the possible introduction of error in the form of distortion, or bias, but also reduce error by reducing the dispersion of awards. In both levels of scenario, the beneficial effect on dispersion dominates any distortionary effect on the size of the award, causing an improvement in accuracy. In both levels of scenario, there is a significant positive effect on accuracy overall—comparing all active treatment conditions combined (i.e., any treatment combination that involves any prior-award information) to the control condition of no prior-award information ($p<0.000^{***}$ for punitive damages and $p=0.008^{**}$ for pain and suffering). Furthermore, prior-award information improves accuracy whether the prior-award information is unbiased low variability ($p<0.000^{***}$ for punitive damages and $p=0.008^{**}$ for pain and suffering), unbiased high variability ($p=0.001^{**}$ for punitive damages and $p=0.013^{*}$ for pain and suffering), biased low variability ($p<0.000^{***}$ for punitive
damages and \( p < 0.000^{***} \) for pain and suffering), or biased high variability (\( p < 0.000^{***} \) for punitive damages and \( p = 0.001^{**} \) for pain and suffering).\(^{106}\)

Moreover, to confirm that our results hold without the influence of the more-extreme values, we examine four important effects using the punitive damages and pain and suffering data truncated at the 90\(^{th}\) percentile—i.e., truncated at $5 million for punitive damages and $10 million for pain and suffering. Using these data, we nevertheless observe a significant improvement in accuracy caused by unbiased low variability (\( p < 0.000^{***} \) for punitive damages and \( p < 0.000^{***} \) for pain and suffering) and biased low variability (\( p < 0.000^{***} \) for punitive damages and \( p < 0.000^{***} \) for pain and suffering) prior-award information for both levels of scenario. Our results for these effects are summarized in Table D and Figure V.3 below.

\(^{106}\) We analyze the effect of form on accuracy as well. The data provide evidence that prior-award information generally improves accuracy regardless of form—i.e., regardless whether such information is presented as an average, list, or range of prior awards.
Table D. Effects on Error When Truncating at the 90th Percentile

<table>
<thead>
<tr>
<th>Effect</th>
<th>Punitive Damages (truncated at $5 million)</th>
<th>Pain and Suffering (truncated at $10 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE (trillions of dollars)</td>
<td>Difference from control (MSE = 4.29) (trillions of dollars) (95% Fisher Interval)</td>
</tr>
<tr>
<td>Unbiased low variability prior-award information</td>
<td>1.39</td>
<td>-2.89 (-3.7, -2.1)</td>
</tr>
<tr>
<td>Biased low variability prior-award information</td>
<td>1.37</td>
<td>-2.92 (-3.8, -2.2)</td>
</tr>
</tbody>
</table>
Figure V.3. Observed mean squared error (MSE) (in trillions of dollars) for unbiased low variability and biased low variability conditions for punitive damages data (left) and pain and suffering data (right), when truncated at the 90\textsuperscript{th} percentile in each scenario. Dotted lines show MSE for control group, and stars indicate statistical significance (after correction for multiple comparisons) of the difference in MSE between each treatment combination and control.

Our results thus provide strong empirical support for the conclusion in The Logic of CCG, that prior-award information improves accuracy under various conditions of bias, variability, and form.

4.3 Effect of Prior-Award Information on Spread

The effects of prior-award information on accuracy evidence effects on the dispersion of awards as well. Specifically, because accuracy is defined using MSE, which can be separated into bias and variance, and because prior-award information cannot reduce bias (or distortion), we know that any improvement in accuracy is due to a reduction in variance. However, we separately test the effect of prior-award information on spread, which we define differently, using the IQR rather than variance. Understanding the effect of
prior-award information on spread, in addition to its effects 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 percentile and the 25th percentiles—that may be obscured in other measures of random variation.

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 indicates that, overall, prior-award information causes a significant reduction in spread ($p<0.000***$). Furthermore, we find that both unbiased prior-award information and biased prior-award information cause a reduction in spread ($p=0.02*$ for unbiased and $p<0.000***$ for biased).\(^\text{107}\)

In the pain and suffering scenario, we detect a significant reduction in spread caused by biased prior-award information ($p<0.000***$), but no significant effect caused by unbiased prior-award information. This latter result is not very surprising, because the range of the prior awards in the pain and suffering scenario—and particularly the high

\(^{107}\) Note, we find that the average level of form tends 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 seem to have “interpreted” (likely implicitly rather than actively) the information as reflecting awards of lower variability.
variability prior-award information—is far larger than the control group IQR.\textsuperscript{108} This causes spread to remain unchanged (\textit{i.e.}, without significant effect), and even to increase in response to high variability prior-award information, notwithstanding an overall reduction in random variation, as reflected in the effects of unbiased prior-award information on accuracy in the pain and suffering scenario.\textsuperscript{109} Effects on spread within each level of scenario are summarized in Table E below. Fisher intervals are provided in Table E for the difference in spread between each treatment level and control, where the control group for punitive damages has a spread of $2,900,000 and the control group for pain and suffering has a spread of $4,500,000. Table F summarizes support from the data for the explanation above, and in note 109, reflecting our analysis regarding the relationship between 1) the

\textsuperscript{108} The reasons for this are twofold: 1) the percentiles chosen for determining the range of values of unbiased high variability prior awards are the 15\textsuperscript{th} and 85\textsuperscript{th} percentiles, substantially wider than the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles; and 2) there is greater spread in the data or sampling variation resulting in greater dispersion in the pilot study awards.

\textsuperscript{109} Unbiased high variability and unbiased low variability prior-awards in the pain and suffering scenario range from $200,000 to $15 million and from $1.12 million to $4.7 million, respectively, compared to the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles of the pain and suffering control group awards, which are $50,000 and $5 million, respectively. This explanation is corroborated by our results in the punitive damages scenario, where, although spread decreases in response to unbiased and biased prior-award information (separately and combined), and decreases in response to unbiased low variability, biased low variability, and biased high variability prior-award information, it increases 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 range in value from $10,000 to $10 million, far greater than the control group 25\textsuperscript{th} and 75\textsuperscript{th} percentiles, $100,000 and $3 million, respectively. (Compare this to the unbiased low variability prior awards, which range only from $500,000 to $1 million, well within the range of control group 25\textsuperscript{th} and 75\textsuperscript{th} percentiles.) This explanation is further supported by previous findings that informing jurors of a damages cap may cause an upward effect on award amounts and dispersion. See Saks, et al., supra note 20, at 249-53. Note that such an effect is not observed for biased prior-award information, since the awards in those treatment conditions are far less variable (and have far lower ranges) than the awards in the unbiased treatment conditions. We thus observe a significant reduction in overall spread caused by biased prior-award information in both levels of scenario (even despite the absence of a detected reduction in magnitude in the pain and suffering scenario (discussed in further detail infra) in response to biased prior-award information).
range of the prior awards provided to participants relative to the control group IQR, and 2) the effect of the prior-award information on spread.
<table>
<thead>
<tr>
<th>Effect</th>
<th>Punitive Damages</th>
<th></th>
<th>Pain and Suffering</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spread (IQR)</td>
<td>Difference from control (IQR = 2,900,000) (95% Fisher Interval)</td>
<td>Unadjust ed p-value</td>
<td>Spread (IQR)</td>
<td>Difference from control (IQR = 4,500,000) (95% Fisher Interval)</td>
</tr>
<tr>
<td>Treatment</td>
<td>$1,430,000</td>
<td>-$1,470,000 (-$1,620,000, -$1,070,000)</td>
<td>0.000***</td>
<td>$4,978,000</td>
<td>$478,000 (-$610,000, $2,400,000)</td>
</tr>
<tr>
<td>Unbiased prior-award information</td>
<td>$2,351,000</td>
<td>-$549,000 (-$749,000, -$70,000)</td>
<td>0.018*</td>
<td>$6,000,000</td>
<td>$1,500,000 (-$72,000, $3,510,000)</td>
</tr>
<tr>
<td>Unbiased low variability prior-award information</td>
<td>$427,500</td>
<td>-$2,472,500 (-$2,720,000, -$1,980,000)</td>
<td>0.000***</td>
<td>$2,750,000</td>
<td>$-1,750,000 (-$3,250,000, $320,000)</td>
</tr>
<tr>
<td>Unbiased high variability prior-award information</td>
<td>$4,825,000</td>
<td>$1,925,000 ($1,675,000, $2,420,000)</td>
<td>0.000***</td>
<td>$10,750,000</td>
<td>$6,250,000 ($4,750,000, $8,310,000)</td>
</tr>
<tr>
<td>Biased prior-award information</td>
<td>$509,000</td>
<td>-$2,391,000 (-$2,580,000, -$1,920,000)</td>
<td>0.000***</td>
<td>$3,955,000</td>
<td>-$545,000 (-$1,650,000, $1,430,000)</td>
</tr>
<tr>
<td>Biased low variability prior-award information</td>
<td>$172,500</td>
<td>-$2,727,500 (-$2,970,000, -$2,280,000)</td>
<td>0.000***</td>
<td>$1,887,500</td>
<td>-$2,612,500 (-$4,110,000, -$566,000)</td>
</tr>
<tr>
<td>Biased high variability prior-award information</td>
<td>$900,000</td>
<td>-$2,000,000 (-$2,280,000, -$1,490,000)</td>
<td>0.000***</td>
<td>$6,000,000</td>
<td>$1,500,000 ($0, $3,560,000)</td>
</tr>
</tbody>
</table>
To confirm our interpretation of our findings regarding spread, and to test its robustness to alternative truncation thresholds and measures of dispersion, we test the effects of unbiased prior-award information on the standard deviation of award determinations using the data for each scenario truncated at the 90th percentile. These tests provide substantial support for our findings and interpretation above. Specifically, in both levels of scenario, we observe that unbiased prior-award information significantly reduces the standard deviation of awards relative to control (\(p<0.000***\) for punitive damages and

---

10 Calculated as length of range divided by length of IQR.
Furthermore, in both levels of scenario, we observe that unbiased low variability prior-award information significantly reduces standard deviation relative to control (p<0.000*** for punitive damages and p<0.000*** for pain and suffering), whereas unbiased high variability prior award information has no significant effect on standard deviation relative to control (as expected, due to the relatively-large dispersion of unbiased high variability prior awards).

One implication of our results with respect to spread is that, although courts need not be concerned that prior-award information will reduce overall dispersion and accuracy, courts should be cognizant of the possibility that providing prior awards that have very high variability may, under certain conditions, increase spread—even while reducing dispersion overall—and thus have a mitigating effect on the benefits of prior-award information for accuracy.

Significantly, in addition to the effects on spread, these tests confirm our belief regarding the robustness of the effect of prior-award information on accuracy. Specifically, notwithstanding any detrimental consequences of the sizeable dispersion of the high variability prior awards (and unbiased high variability prior awards in particular), we observe that prior-award information has a significantly positive effect on accuracy across the board.

4.4 Effect of Prior-Award Information on Magnitude

Our interest is in testing whether unbiased and biased prior-award information (as well as interactions with the variability of prior awards) cause any distortion (or “bias”) in the size of awards. Pursuant to our Analysis Protocol, we begin by examining the effect on
magnitude, using the mean of the log-transformed data for each scenario and within levels of bias and variability. We observe, however, that this measure may result in somewhat misleading results for the current study.

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 data points more than it “pulls in” moderate data points. Under certain conditions, however, using the mean of the log-transformed data can distort the sign of an observed effect. This can occur, as it did in the current study—and, likely, because participants chose values from their imagination—when some awards are sufficiently extreme that the observed effect of the non-transformed data is dominated by the effect of the extreme awards. Therefore, when those awards are “pulled in” more substantially than more moderate awards, the observed effect is reversed or otherwise distorted.

Now, arguably, the mean of the log-transformed data is a more appropriate measure with which to test magnitude, since it mitigates the otherwise-overwhelming influence of outliers. In this study, however, this is likely not the case. Specifically, although we observe an increase in the magnitude of awards resulting from unbiased prior-award information (for both pain and suffering and punitive damages), most tests involving other measures of award size result in no significant effect. For example, using the mean results in no significant effect of unbiased prior-award information on award size in the pain and suffering scenario, and a negative effect of unbiased prior-award information on award size in the punitive damages scenario (and no effect if the punitive damages data is truncated at
the 99th percentile). It is possible that these effects are heavily influenced by extreme values; but, arguably, the influence of extreme values is significant in this study—particularly where the datasets are truncated at the 99th percentile, since these datasets already exclude the most extreme data points.

Furthermore, for both punitive damages and pain and suffering, we observed no differences between the median of the control group awards and the median of the awards in conditions involving unbiased prior-award information. As opposed to the mean of the unlogged data, this measure excludes the effect of extreme values. We also tested the effect of unbiased prior-award information on award size using as a measure of size the mean of awards in the data truncated at the 90th percentile, finding no significant effect for punitive damages and a positive effect for pain and suffering. We summarize the effects of unbiased prior-award information on award size under alternative measurements in Table G below.

Table G. Unbiased vs. Control: Alternative Measurements and Corresponding Effects on Award Size

<table>
<thead>
<tr>
<th></th>
<th>Punitive Damages</th>
<th>Pain and Suffering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log mean</td>
<td>Median</td>
</tr>
<tr>
<td>Unbiased versus control</td>
<td>Positive</td>
<td>0</td>
</tr>
</tbody>
</table>

Similarly, for biased prior-award information in the pain and suffering scenario, using either the mean or median of the unlogged data reveals no significant effect on
magnitude, relative to awards in the control condition, whereas using the mean of the logged data indicates a positive effect, relative to awards in the control condition (p=0.02*).

Thus, our results regarding the effect of unbiased prior-award information with respect to award size can be described as inconclusive. Although, using the mean of the log-transformed data, we detect a positive effect of unbiased prior-award information on magnitude (p<0.000*** for punitive damages and p<0.000*** for pain and suffering), at best, the tests are very sensitive to the measure used, and, at worst, the log-transform causes dramatic distortions. In any event, using other common measures of size, such as the median, we detect no significant difference between the size of awards in the control condition and the size of awards in the unbiased conditions.

Of course, we can attempt to interpret this breakdown of effects. One way of understanding it is as follows: unbiased prior-award information has no significant effect on the size of awards, in the sense that the central award is not affected. For example, in the punitive damages scenario, the median of awards in both the control condition and across the unbiased conditions is $1 million. On the other hand, unbiased prior-award information has no effect or a significant negative effect on award size when the effects of extreme awards are considered. This is corroborated by our exploratory analysis regarding outliers,111 and is significant in the current study, since extreme awards, as well as random variation generally, is the problem that prior-award information seeks to address. Finally, unbiased prior-award information may cause greater award size, in the sense that the

---

111 See infra § 4.6.
awards in the *unbiased* conditions are greater when the effect of more extreme awards is excluded but the effect of some level of extremeness is included. One possible explanation for this effect is that the observed detrimental effect of the sizeable dispersion of *high variability* prior-award information on *spread* may also cause some positive effect on award size, due to the lopsided nature of the *high variability* prior awards, which are bounded from below at $0$ but unbounded from above and right skewed. In any event, this latter effect is interesting, but not very robust, since the results revert to insignificance when we use (unlogged) data truncated at the 99th percentile to exclude extreme outliers, and mixed results when we use (unlogged) data truncated at the 90th percentile.

Additionally, in line with our expectations, for *biased* prior-award information in the *punitive damages scenario*, the data evidence a significant negative effect on *magnitude* (measured using the mean of the *logged data*) ($p<0.000^{***}$). This result is consistent with the effect direction observed using other measures of award size, such as the *median* or *mean*. Further, for *punitive damages*, the overall effect of prior-award information on *magnitude*, relative to awards in the control condition, is not significant. Similarly, we observe no significant effect on award size when measured using the *mean* of the *logged data* or the *median*, and a downward effect (in line with our expectations, due to the effect of *biased* prior-award information) when award size is measured using the *mean* of the raw data ($p<0.000^{***}$).

Finally, regardless whether prior-award information affects award size, in all levels of *bias* and *variability*, and for both levels of *scenario*, we observe that prior-award information improves *accuracy*. 
4.5 Effect of Low Variability Prior-Award Information Relative to High Variability Prior-Award Information

We find strong evidence, in line with our expectations, that low variability prior-award information is more influential than high variability prior-award information.

First, for both levels of scenario, the data indicate that low variability prior-award information causes a reduction in spread ($p<0.000^{***}$ for punitive damages and $p<0.000^{***}$ for pain and suffering) and a reduction in magnitude ($p<0.000^{***}$ for punitive damages and $p<0.000^{***}$ for pain and suffering), relative to 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,\textsuperscript{112} and thus has a greater influence on award determinations. This “influence” translates to greater impacts on spread and magnitude.

Thus, for both levels of scenario, unbiased low variability prior-award information, relative to unbiased high variability prior-award information, has no significant effect on magnitude and a significant negative effect on spread ($p<0.000^{***}$ for pain and suffering and $p<0.0000^{***}$ for punitive damages). These results are in line with our expectations: unbiased prior-award information is expected to have no significant effect on magnitude, whether it is low variability or high variability, and prior-award information is expected to have a greater negative impact on spread if it is low variability than if it is high variability.

\textsuperscript{112} See supra § 2.3.
Moreover, for both levels of scenario, biased low variability prior-award information, relative to biased high variability prior-award information, has a significant negative effect on both magnitude ($p<0.000***$ for pain and suffering and $p<0.000***$ for punitive damages) and spread ($p<0.000***$ for pain and suffering and $p=0.003**$ for punitive damages). These results are also in line with our expectations. Specifically, because the prior-award information is biased, there are two effects—low variability prior-award information has greater influence than high variability prior-award information, and therefore both a greater distortionary, or “biasing,” effect on magnitude (here, a negative effect), and a greater “narrowing” effect on spread. Our results regarding the effects of low variability prior-award information relative to high variability prior-award information are summarized in Table H below.

<table>
<thead>
<tr>
<th>Table H. Low Variability vs. High Variability Prior-Award Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
</tr>
<tr>
<td><strong>Low variability vs. high variability</strong></td>
</tr>
<tr>
<td>MSE (in trillions)</td>
</tr>
<tr>
<td>Spread</td>
</tr>
<tr>
<td>Magnitude</td>
</tr>
<tr>
<td><strong>Low variability unbiased vs. high variability unbiased</strong></td>
</tr>
<tr>
<td>MSE (in trillions)</td>
</tr>
<tr>
<td>Spread</td>
</tr>
<tr>
<td>Magnitude</td>
</tr>
<tr>
<td><strong>Low variability biased vs. high variability biased</strong></td>
</tr>
<tr>
<td>MSE (in trillions)</td>
</tr>
<tr>
<td>Spread</td>
</tr>
<tr>
<td>Magnitude</td>
</tr>
</tbody>
</table>
Note, 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 discussed above, it is possible that the high variability prior-award information causes an observed increase in spread, since the range of the high variability prior awards is greater than the IQR. It is therefore possible that the observed effect here results from removing high variability awards and thus removing the upward effect on spread. Indeed, such an effect could also explain an observed decrease in spread caused by unbiased low variability prior-award information relative to unbiased high variability prior-award information. In other words, it is possible to argue that the observed reduction in spread has nothing to do with the influence of low variability, versus high variability, prior-award information, but rather, that it is caused merely by removing the upward effect on spread caused by the sizeable dispersion of the high variability prior awards.

It is more difficult, however, to apply this explanation to the effects of biased low variability prior-award information, relative to biased high variability prior-award information, with respect to spread and magnitude. In contrast to unbiased prior-award information, it is unlikely that either biased low variability prior-award information or biased high variability prior-award information would cause an increase in spread, relative to the control condition. Indeed, in both levels of scenario, neither biased low variability prior-award information nor biased high variability prior-award information has such an effect. This is because the variability of such information is reduced substantially by the
bias ratio, leading to muted levels of variability for low variability and high variability prior-award information in the biased conditions relative to their counterparts in the unbiased conditions.

Furthermore, remember that biased low variability prior-award information does not involve prior awards that are of lower amount than biased high variability prior-award information; rather, it only has lower variability. The biased low variability prior-award information (ranging from $50,000 to $100,000) is completely contained within the range of values in the biased high variability prior-award information (ranging from $1,000 to $1 million).

Thus, the observed significant negative effects of biased low variability prior-award information, relative to biased high variability prior-award information, on spread and magnitude, and within both levels of scenario, provide strong evidence that low variability prior-award information has greater influence—i.e., is more impactful—on award determinations than high variability prior-award information. This finding provides support for the model assumption in The Logic of CCG, that the influence of prior-award information is proportional to the inverse of the variability of the awards.

4.6 The Effect of Prior-Award Information on Outliers

In our exploratory analysis, we examine the effect of prior-award information on “outliers,” which we define as awards at or above the 99th percentile ($50 million in the punitive damages scenario and $100 million in the pain and suffering scenario).

\footnote{113} See supra § 3.
In the control group of the *punitive damages scenario*, there are thirteen observations out of 249 (5.2%) above $50 million, whereas in the ten treatment arms of the *punitive damages scenario*, there are 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 \(\chi^2_i = 29.2, \ p<0.000^{**}\). Furthermore, in the five *unbiased* treatment arms in the *punitive damages scenario*, there are 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 \(\chi^2_i = 17.6, \ p<0.000^{***}\).

In the control group of the *untruncated* dataset for the *pain and suffering scenario*, there are ten observations out of 244 (4.1%) above $100 million, whereas in the ten treatment arms of the *pain and suffering scenario*, there are 33 observations out of 2463 (1.3%) above $100 million, or, on average, approximately three observations \((33/10 = 3)\) above $100 million per 249 observations. This represents a significant reduction in outliers \(\chi^2_i = 9.11, \ p=0.002^{**}\). Furthermore, in the five *unbiased* treatment arms in the *pain and suffering scenario*, there are 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 \(\chi^2_i = 5.6, \ p=0.018^*\).

Our analysis regarding the effect of prior-award information on outliers is summarized in Table I below.
Table I. Absolute Differences in Proportions of Outliers

<table>
<thead>
<tr>
<th></th>
<th>Punitive Damages (Raw)</th>
<th>Pain and Suffering (Raw)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment rate</td>
<td>Control rate</td>
<td>Unadjusted p-value</td>
<td>Treatment rate</td>
<td>Control rate</td>
<td>Unadjusted p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment vs. Control</td>
<td>0.9%</td>
<td>5.2%</td>
<td>0.000***</td>
<td>1.3%</td>
<td>4.1%</td>
<td>0.002**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbiased vs. Control</td>
<td>1.1%</td>
<td>5.2%</td>
<td>0.000***</td>
<td>1.6%</td>
<td>4.1%</td>
<td>0.018*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

The “stark unpredictability” of awards for pain and suffering and punitive damages is arguably unacceptable.\(^{114}\) We desire “predictability and proportionality,”\(^{115}\) but, at the same time, require juries to determine these awards through nothing more than a “standardless, unguided exercise of discretion.”\(^{116}\) 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.

*The Logic of CCG* addressed major objections to the use of prior-award information to guide award determinations, and argued that CCG not only reduces the variability of awards, but improves accuracy—that is, bias and variance—generally. In particular, the paper relied on certain behavioral assumptions to explain why any introduction of bias caused by CCG would be well outweighed by a reduction in variability; and that such effects

---


\(^{115}\) Payne v. Jones, 711 F.3d 85, 94 (2d Cir. 2013).

\(^{116}\) Jutzi-Johnson v. United States, 263 F.3d 753, 759 (7th Cir. 2001).
would be sufficiently predictable that, absent extraordinary circumstances, courts generally need not be concerned regarding whether or not CCG will improve accuracy in a particular case.\textsuperscript{117} It concluded that, “[i]n short, we can expect that a reasonable method for identifying prior ‘comparable’ cases will result in prior-award information that is likely to improve [accuracy, and thus,] reliability.”\textsuperscript{118} Furthermore, a second paper derived various mathematical results demonstrating the robustness of CCG methods to a wide range of conditions, including those involving the identification of “misaligned” prior cases.\textsuperscript{119} But these results rely on the same behavioral assumptions as those relied upon in \textit{The Logic of CCG}.

The purpose of the current study is to examine whether the behavioral assumptions relied upon in these papers are justified—that is, whether they are supported by the data. In particular, we are interested in testing, behaviorally, whether prior-award information improves \textit{accuracy} under a robust set of conditions, and whether jurors behave as the model predicts in response to \textit{bias} and \textit{variability} conditions, and prior-award information generally.

In the present study, we find that prior-award information improves \textit{accuracy}, not only when such information is \textit{unbiased}, but also when it is \textit{biased}, and even \textit{biased and low variability}. That is, even when the prior awards are narrowly distributed around a “misaligned” center—one that is substantially less than the “correct” center—the beneficial

\begin{footnotesize}
\begin{enumerate}
\item See \textit{The Logic of CCG} § 4.
\item \textit{Id.} at 21.
\item See Bavli & Chen, supra note 50.
\end{enumerate}
\end{footnotesize}
effects of prior-award information with respect to random variation, or *judgment variability*, are greater than the detrimental effects with respect to the expected size of the award. Thus, any *error* caused by the distortion of award size, or the introduction of bias, is outweighed by the reduction in *error* caused by reducing dispersion.

Indeed, this is confirmed, and given additional color, by our *other* findings. First, it is clear from our descriptive statistics that awards in the *punitive damages scenario* and the *pain and suffering scenario* are extremely unpredictable. Second, we find that, in addition to its effect on random variation generally, prior-award information has a significant downward effect on *spread*. We also find that it has a significant downward effect on *outliers* in particular. Thus, prior-award information seems to have beneficial effects on *judgment variability* at a general level—it reduces *spread*, the frequency of outliers, and variance generally. Note, however, that a number of findings regarding *spread* indicate that, if the prior awards are sufficiently variable, they may cause some subset of awards to increase in variation, even while the random variation of awards may decrease overall. Third, although we do not arrive at a conclusive finding regarding the effect of *unbiased* prior-award information on *magnitude*—*e.g.*, we are unable to confirm the result of Saks et al. that *unbiased* prior-award information causes no change in the amounts awarded and therefore no introduction of *error* in the form of bias\(^{120}\)—the data provide evidence that *biased* prior-award information has a significant downward effect on *magnitude*, as predicted. Fourth, in line with the behavioral assumptions in *The Logic of CCG*, we find that

---

\(^{120}\) Saks et al, *supra* note 20 at 249–53.
*low variability* prior-award information has significantly greater influence on award determinations than *high variability* prior-award information, and therefore significantly greater impact on *spread* and *magnitude*.

Note, however, that the effect of prior-award *variability* on *spread* and *magnitude* may interact with the effect described above, whereby *high variability* prior awards may increase the *spread* of certain award determinations. Specifically, it is possible that, for highly variable prior awards, two effects may interact: 1) the highly variable prior awards have a reduced influence on award determinations (relative to less variable prior awards), resulting in a muted effect on *spread* and *magnitude*; and 2) the prior awards may have an upward impact on *spread* and *magnitude*. It is unclear which effect would dominate in a particular application of CCG; but our findings suggest that the latter effect would, at most, be concerning when the prior awards are extremely variable—and even then, as evidenced by the data, the prior awards are likely to improve accuracy on balance.

Although we uncover substantial evidence in support of our hypotheses and the argument in *The Logic of CCG*, there are a number of important limitations to our study. First, as highlighted above, some results give rise to the possibility of effects that are not described by our hypotheses, or that we explain as effects related to measurement. Although these issues are generally unsurprising, and our explanations perhaps uncontroversial, conclusive reasoning may require additional examination.

Second, the experimental units in this study are mock *jurors* rather than mock *juries*, whereas *juries* are the units at issue in the context of real-world award determinations. Of course, *juries* are composed of *jurors*, and the behavior of *jurors* are likely closely tied to
that of *juries*; nevertheless, extending conclusions from the present study to *juries* requires additional inference that should be considered carefully. For example, it is likely that the random variation of awards is less for *juries* than *jurors*, where the former deliberate and combine—in some sense, compromise over—heterogeneous views regarding an appropriate award.\(^{121}\)

Third, our experimental units are *mock* jurors, rather than real-world jurors who decide an award following an actual trial (rather than a description). In a sense, this study is a simplified “laboratory” experiment aimed at studying juror behavior. A more ideal (although far costlier) experiment would involve an intervention in real-world award determinations using *juries* rather than *jurors*, and *real-world* *juries* in the context of *real-world trials*. For example, the variability of awards may be exaggerated, because mock jurors treat the situation as hypothetical and ignore certain real consequences of extreme awards, such as bankruptcies, job loss, etc.\(^ {122}\) Note that extreme awards are observed in the real world, but likely less frequently (and perhaps less dramatically) than in survey experiments. Additionally, the summary description provided to participants in each level of *scenario* is likely to affect the variability of awards (both in control and active treatment conditions) relative to real-world trials in which multifaceted evidentiary support is provided to substantiate arguments by the plaintiff and the defendant, and presided over by a judge.

\(^{121}\) See Diamond et al., *supra* note 7, at 316 (comparing juror and jury damage awards); Sanders, *supra* note 13, at 494-96 (discussing criticisms based on the disparity between *juror*-based studies and *jury* awards).

\(^{122}\) 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.

226
Fourth, we used Mechanical Turk to administer the study. Although we compared the population of Mechanical Turk users to the population of individuals on U.S. juries, and concluded that these populations are similar based on the covariates examined, there are likely to be some differences that may affect our results.123 Additionally, although we attempted to exclude individuals not qualified for participation on a U.S. jury, neither our technical inclusion criteria nor our stated inclusion criteria ensured compliance, and we did not expect to exclude 100% of individuals ineligible for jury service. To the extent that certain participants should have been excluded from the study based on their ineligibility to serve on a U.S. jury, and to the extent that their responses to our survey differ from the responses we would receive from the true population of eligible jurors, such participation may affect our results.

Notwithstanding these limitations, and others, our findings provide strong evidence that CCG improves accuracy under a robust set of conditions, and that the behavioral assumptions and conclusions in The Logic of CCG are supported by the data. In summary, our research provides substantial support for the argument that CCG is an effective method for reducing the unpredictability 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.