2000

Deliberating about Dollars: The Severity Shift Empirical Study

Cass R. Sunstein

Daniel Kahneman

David Schkade

EMPIRICAL STUDY

DELIBERATING ABOUT DOLLARS: THE SEVERITY SHIFT

*David Schkade,**
*Cass R. Sunstein,***
*and Daniel Kahneman****

How does jury deliberation affect the predeliberation judgments of individual jurors? In this paper we make progress on that question by reporting the results of a study of over 500 mock juries composed of over 3000 jury eligible citizens. Our principal finding is that with respect to dollars, deliberation produces a "severity shift," in which the jury's dollar verdict is systematically higher than that of the median of its jurors' predeliberation judgments. A "deliberation shift analysis" is introduced to measure the effect of deliberation. The severity shift is attributed to a "rhetorical asymmetry," in which arguments for higher awards are more persuasive than arguments for lower awards. When judgments are measured not in terms of dollars but on a rating scale of punishment severity, deliberation increased high ratings and decreased low ratings. We also find that deliberation does not alleviate the problem of erratic and unpredictable individual dollar awards, but in fact exacerbates it. Implications for punitive damage awards and deliberation generally are discussed.

INTRODUCTION

How, if at all, is the outcome of group deliberation different from a statistical aggregation of individual predeliberation judgments? How might jury deliberations depart from the median or mean of individual judgments? Speculation is not difficult. Perhaps juries converge toward the midpoint of individual judgments; perhaps juries move away from, or toward, the high or low of individual extremes. Perhaps juries produce an outcome that is more just or more accurate; perhaps juries generate more predictable and less erratic judgments, so that unpredictability at the individual level, or at the level of the mean or median of (six or twelve) individual judgments, is further reduced by deliberation at the

* Herbert D. Kelleher/MCorp Professor of Business, Graduate School of Business, University of Texas, Austin.

** Karl N. Llewellyn Distinguished Service Professor of Jurisprudence, Law School and Department of Political Science, University of Chicago.

*** Eugene Higgins Professor of Psychology and Professor of Public Affairs, Princeton University. The authors are grateful to Exxon Company, U.S.A. for support of the research in this Empirical Study. Exxon bears no responsibility for our analysis or our conclusions; the data reported and the opinions expressed here belong to the authors. For helpful comments we are grateful to participants in workshops at Harvard, Stanford, and the University of Chicago, and also to Robert MacCoun, Eric Posner, Richard Posner, Jeffrey Rachlinski, and Michael Saks.

jury level. A pervasive question is whether a deliberating jury has the effect of producing outcomes that treat the similarly situated similarly—perhaps in terms of civil or criminal liability (do people who have engaged in the same conduct receive the same verdict?), perhaps in the determination of appropriate damage awards, either compensatory or punitive (do similarly situated people receive the same awards?).[1]

In this Empirical Study, we attempt to make some progress on these questions. We do so principally by reporting the results of an extensive study of mock juries (over 3000 people and 500 juries in total). Six-person juries were asked to deliberate about the appropriate punishment in civil cases involving personal injury in two ways, by setting punitive awards in dollars and by indicating, on a rating scale, the severity of the punishment they wished to inflict on the defendant. Our most important and general finding is that with respect to dollar awards, deliberation produces a *severity shift*: The jury's dollar verdict is typically higher, and often far higher, than the median judgment of the same jury's individual members before deliberation began.[2]

To compress a long story, our specific findings are these:

- Jurors followed a simple principle of majority rule in deciding whether to impose punitive damages at all; the decision to award damages was largely a function of the majority of individual predeliberation votes.
- Where the median of individual predeliberation judgments favored a *high* punishment rating, deliberation typically *increased* the rating of the group.
- Where the median of individual predeliberation judgments favored a *low* punishment rating, deliberation typically *decreased* the rating of the group.
- Where the median of individual predeliberation judgments favored *large* dollar awards, deliberation typically *increased* the dollar award of the group, often dramatically so: *Among juries that voted to award punitive damages, 27% reached dollar verdicts that were as high as or higher than the highest predeliberation judgment of their individual members.*
- Where the median of individual predeliberation judgments favored small dollar awards, deliberation typically *increased*

1. For concerns along this line, see BMW of North America, Inc. v. Gore, 517 U.S. 559, 587 (1996) (Breyer, J., concurring) ("Requiring the application of law, rather than a decisionmaker's caprice, does more than simply provide citizens notice of what actions may subject them to punishment; it also helps to assure the uniform general treatment of similarly situated persons that is the essence of law itself.").

2. Compare the finding of a "leniency shift" in criminal juries and that this shift produces more accurate judgments, because juries are more likely than individual jurors to apply the reasonable doubt standard correctly. See Robert J. MacCoun & Norbert L. Kerr, Asymmetric Influence in Mock Jury Deliberation: Jurors' Bias for Leniency, 54 J. Personality & Soc. Psychol. 21, 21–22 (1988) (reviewing literature and empirical findings regarding hypothesis that deliberation makes jurors more lenient).

the dollar award of the group, though the increase was smaller than for high dollar awards.

- With respect to punishment ratings, juries were neither more nor less consistent and predictable than the mean or median juror. With respect to dollar awards, juries were less consistent and predictable than the mean or median juror. *With respect to dollar awards, jury deliberation substantially increased unpredictability.*

For punishment ratings, the principal effect of deliberation was thus to move group judgments toward a more extreme version of the original tendency (low or high) of individuals within that group. This effect is, we believe, closely related to phenomena frequently studied under the labels of "risky shifts," "choice shifts," and "group polarization."[3] For dollar awards—the more important issue—the effect of deliberation was to produce a severity shift, such that juries' dollar verdicts were systematically higher than the median predeliberation judgments of jurors. The severity shift stems, we believe, from a systematic *rhetorical advantage* held by those arguing for higher dollar awards, an advantage that operates independently of the particular case at issue.

The study reported here has the advantage of being extremely close to—in fact part of the design is based on—an earlier one involving not deliberating juries but responses of 899 individuals to punitive damage cases.[4] Our earlier study focused on the question of predictability, which we understood to be a function of whether the judgment of one randomly selected jury is a good predictor of the judgment of other randomly selected juries judging the same case.[5] We found a remarkable consensus in the judgments of individual jurors, made on a rating scale, about a series of personal injury cases. That study therefore found that *with respect to the underlying moral evaluation, groups of different (non-deliberat-*

---

3. See, e.g., Daniel J. Isenberg, Group Polarization: A Critical Review and Meta-Analysis, 50 J. Personality & Soc. Psychol. 1141 (1986); Craig McGarty et al., Group Polarization as Conformity to the Prototypical Group Member, 31 Brit. J. Soc. Psychol. 1, 3 (1992); David G. Myers & Helmut Lamm, The Group Polarization Phenomenon, 83 Psychol. Bull. 602 (1976); Russell Spears et al., De-Individuation and Group Polarization in Computer-Mediated Communication, 29 Brit. J. Soc. Psychol. 121 (1990). More particularly, our data show a choice shift. See Johannes A. Zuber et al., Choice Shift and Group Polarization: An Analysis of the Status of Arguments and Social Decision Schemes, 62 J. Personality and Soc. Psychol. 50, 50 (1992) (defining "choice shift" as "the difference between the arithmetic mean of the individual first preferences before discussion . . . and the group decision").

4. See Cass R. Sunstein, Daniel Kahneman, & David Schkade, Assessing Punitive Damages (with Notes on Cognition and Valuation in Law), 107 Yale L.J. 2071 (1998). This presentation is geared to analysis of the legal issues; the underlying data, and relevant psychological points, are presented in more detail in Daniel Kahneman, David Schkade, & Cass R. Sunstein, Shared Outrage and Erratic Awards: The Psychology of Punitive Damages, 16 J. Risk & Uncertainty 49 (1998).

5. Thus variance among juries exposed to the same case is our basic measure of unpredictability. We use the term "erratic" as a synonym for unpredictable.

ing) jurors[6] *are likely to reach similar conclusions about the relative severity of different cases.*[7] Thus all-white, all-poor, all-rich, all-educated, all-poorly-educated, all-male, all-female, all-young, and all-old juries would probably come to very similar rankings of a set of cases, at least in personal injury cases and very possibly elsewhere.[8] It follows that the median of the individual judgments of any random group of twelve people is likely to produce a moral judgment that predicts, with a reasonably high degree of accuracy, the judgment of any other group of twelve people (also defined by the median of the individual judgments of group members).

At the same time, the study found that *assessment of cases in terms of dollars produces great unpredictability.*[9] To be sure, ranking the cases by their aggregate dollar awards or by their aggregate punishment ratings produced very similar orderings of the cases from least to most severe. But dollar awards are unpredictable in the specific sense that punishment ratings are not: The judgment of any particular group of twelve (determined by taking the median judgment as that of the group) is a poor predictor of the judgment of other groups of twelve (determined in the same way).[10] We showed that the same case, presented to different jurors, will elicit similar ratings but quite different dollar awards, producing a situation where the similarly situated are not treated similarly. This unpredictability may well produce overdeterrence in risk-averse defendants or in any case muffled and confusing signals.[11] We concluded that the unbounded dollar scale contributes to evidently erratic monetary judgments in many areas of the law, including not only punitive damages but also compensatory awards in cases involving libel, sexual harassment,[12]

---

6. As explained below, see infra text accompanying notes 15–16, we looked at "statistical juries" consisting of random groups of 12 individual judgments, with the mean or median judgment of each group of 12 reflecting the "verdict." See Sunstein, Kahneman, & Schkade, supra note 4, at 2100–01.

7. See id. at 2095–100.

8. See id. at 2097–100.

9. See id. at 2100–04.

10. Note that we hold constant several factors that can be used to capture some of the variability in punitive damage awards, such as compensatory damages, case category, case particulars, and jurisdiction. It has been proposed by some authors that when analyzed using these factors, punitive awards are reasonably predictable. See Theodore Eisenberg et al., The Predictability of Punitive Damages, 26 J. Legal Stud. 623, 644 (1997). Because we hold these factors constant, they cannot account for the unpredictability that we documented previously and that we document here.

11. See Paul H. Rubin et al., BMW v. Gore: Mitigating the Punitive Economics of Punitive Damages, 5 Sup. Ct. Econ. Rev. 179, 184 (1997).

12. See Judy Shih & Cass R. Sunstein, Damages in Sexual Harassment Cases 1–3, *in* Sexual Harassment (Catharine MacKinnon & Reva Siegel eds., forthcoming 2000) (manuscript on file with the *Columbia Law Review*) (finding that compensatory and punitive damage awards are random in sexual harassment cases).

pain and suffering,[13] and intentional infliction of emotional distress.[14]

Our earlier study did not involve deliberating juries. In the absence of evidence about how deliberation would affect individual judgments, we analyzed statistical juries, by treating the median of a deliberating group as a good predictor of the ultimate judgment of the deliberating jury.[15] We did so on the ground that the received wisdom seemed to support this approach.[16] In this Empirical Study, we investigate the received wisdom—and find overwhelming evidence that it is wrong. The dollar awards of deliberating groups were not close to any measure of central tendency; they were much higher. We also explore several questions of importance to those interested in damages, juries, and deliberative processes in general.[17] The answers have implications not only for punitive awards, but also for other damage judgments (certainly when these are hard to monetize), possibly for questions of civil and criminal liability as well, and even for deliberation generally.

As we have noted, the present study finds that as compared with the median of individual predeliberation judgments, deliberation significantly increases high dollar awards, increases high punishment ratings, decreases low punishment ratings, and modestly increases low dollar awards. To summarize a complex analysis, it follows that *deliberating juries produce even more unpredictability than was observed for statistical juries.* Moreover, dollar responses vary much more across juries than does punitive intent on a rating scale. Thus we find shared moral judgments but erratic dollar awards not only for individuals but for deliberative juries as well.

What follows at the normative level? Without an independent theory of what awards should be, the evidence found here does not show whether deliberation, and the resulting severity shift, make dollar awards better or worse. The safest and most cautious conclusion is that to the extent that unpredictable punitive damage awards raise a serious concern, the problem is not removed by deliberation. To the extent that unpredictability is a problem, our findings about the outcomes of jury deliberation—predictable moral judgments but unpredictable dollar awards—raise further questions about whether punitive awards should be made by juries, or instead by judges or some kind of administrative insti-

---

13. See David W. Leebron, Final Moments: Damages for Pain and Suffering Prior to Death, 64 N.Y.U. L. Rev. 256, 259 (1989) (arguing that tort awards for pain and suffering "vary significantly and that neither the specific facts of the case nor differing theoretical views on the functions of the awards can explain such variation").

14. See Sunstein, Kahneman, & Schkade, supra note 4, at 2131–40.

15. See id. at 2101.

16. See id.

17. On the latter topic, see Amy Gutmann & Dennis Thompson, Democracy and Disagreement 199–230 (1996) (exploring how a deliberative perspective can provide guidance in dealing with moral disagreement in politics); our analysis of group polarization raises some questions for the deliberative conception of democracy, though we leave those questions largely implicit here.

tution.[18] Perhaps juries should be asked to come up with punishment ratings, not with dollar awards, and the legal system should "translate" jury ratings into dollar awards by some preset formula. Our findings also raise a set of novel issues about deliberation as a whole. Is deliberation anything to celebrate if groups tend to move further in the direction suggested by their original tendency—if (for example) high dollar awards go up, low punishment ratings go down, and groups opposed to gun control and in favor of affirmative action end up thinking a more extreme version of what individual group members originally thought? We offer a brief discussion of some of these issues, with particular reference to the effects of deliberation on punishment ratings and on dollar awards.

I. Theoretical Preliminaries: Outrage and Scales

Jury awards of punitive damages have become one of the most controversial topics in modern public law.[19] To take just one example, an award of $4.9 billion against General Motors attracted a great deal of national attention in July of 1999.[20] It is now clear that the Due Process Clause imposes constraints on permissible awards.[21] A number of statutes, enacted and proposed, create punitive damage "caps,"[22] and high awards have become a primary impetus for tort reform in general.[23] There are also controversial issues about punitive damage awards in civil

18. See, e.g., W. Kip Viscusi, Corporate Risk Analysis: A Reckless Act?, 52 Stan. L. Rev. 547, 549–52 (2000) (challenging jury verdicts on ground that they are often irrational). The point is generally discussed in Sunstein, Kahneman, & Schkade, supra note 4, at 2126–30.

19. See, e.g., William M. Landes & Richard A. Posner, The Economic Structure of Tort Law 160–65, 184–85, 223–24 (1987) (developing economic model of damages); Marc Galanter & David Luban, Poetic Justice: Punitive Damages and Legal Pluralism, 42 Am. U. L. Rev. 1393, 1394–96 (1993) (defending punitive damages as a morally necessary part of a private law system); David G. Owen, The Moral Foundations of Punitive Damages, 40 Ala. L. Rev. 705, 705–08 (1989) (analyzing punitive damage rules in terms of moral theory and arguing that moral philosophy supports such rules); A. Mitchell Polinsky & Steven Shavell, Punitive Damages: An Economic Analysis, 111 Harv. L. Rev. 869, 870–76 (1998) (arguing that punitive damages should only be imposed if an injurer has a chance of escaping liability for an injury it caused, and claiming that current legal rules fall short of this goal); Symposium, The Future of Punitive Damages, 1998 Wis. L. Rev. 1, 1–426 (offering various assessments of the "future of punitive damages").

20. See General Motors Appeals Record Lawsuit Damages, N.Y. Times, July 31, 1999, at A9.

21. See BMW of North America, Inc. v. Gore, 517 U.S. 559, 574–75 & n.22 (1996).

22. See generally Developments in the Law—The Civil Jury, 110 Harv. L. Rev. 1408, 1533 & n.158 (1997) (discussing actual and proposed caps). A number of state supreme courts have invalidated such measures. See, e.g., State ex rel. Ohio Academy of Trial Lawyers v. Sheward, 715 N.E.2d 1062, 1090–95 (Ohio 1999) (striking down caps on punitive and general damages as violating state constitutional provisions of right to jury trial and due process).

23. See, e.g., Product Liability Reform Act of 1997, S. 648, 105th Cong. § 2(a)(2) (1997) (listing "[e]xcessive, unpredictable, and often arbitrary damage awards" as a factor motivating the bill). For a general discussion skeptical of the attention paid to high awards, see Marc Galanter, Shadow Play: The Fabled Menace of Punitive Damages, 1998

rights cases, most notably in sexual harassment cases.[24] At the same time, the problems created by punitive awards bear on related questions in other areas of the law, involving, for example, compensatory damages for pain and suffering, libel, and intentional infliction of emotional distress.[25] Similar problems arise in the area of criminal sentencing whenever an administrative agency is asked to impose civil fines.[26]

Participants in the legal system are often asked to come up with some kind of judgment, factual or normative, and then to "translate" that judgment into a dollar award. In the area of punitive damages, it is necessary to make some assessment of the character of the defendant's behavior, and then to ascertain the appropriate dollar amount to be paid to the plaintiff by way of punishment. In many domains, compensatory judgments raise similar puzzles. While juries are nominally expected to find a "fact"—what amount of money would restore the plaintiff to the status quo ante?—it is often extremely difficult to monetize the relevant harm, and normative judgments undoubtedly play a significant role.[27] In the case of punitive damages, it is extremely difficult for even experts to agree on what dollar amount constitutes adequate "punishment" or produces an appropriate deterrent signal.

In all of these areas, the legal system is pervaded by a degree of unpredictability and variance, resulting in apparent arbitrariness, as the similarly situated are treated differently.[28] An extensive study of pain and suffering cases found that as much as 60% of the awards consists of

---

Wis. L. Rev. 1, 5–11 (arguing that the concern over high awards is the result of exaggerations).

24. See Shih & Sunstein, supra note 12 (manuscript at 1–3) (concluding that both compensatory and punitive damage awards were "quite random" in sexual harassment cases).

25. See, e.g., David Baldus et al., Improving Judicial Oversight of Jury Damages Assessments: A Proposal for the Comparative Additur/Remittitur Review of Awards for Nonpecuniary Harms and Punitive Damages, 80 Iowa L. Rev. 1109, 1112–13 (1995) (analyzing common law methods of valuing nonpecuniary harms and discussing additur/remittitur review as a means to control outlying awards); Randall R. Bovbjerg et al., Valuing Life and Limb in Tort: Scheduling "Pain and Suffering," 83 Nw. U. L. Rev. 908, 909 (1989) (proposing three alternative frameworks for valuing non-economic damages); Leebron, supra note 13, at 288–309 (reviewing the factors that cause variability in awards for pain and suffering prior to death).

26. See Edward L. Rubin, Punitive Damages: Reconceptualizing the Runcible Remedies of Common Law, 1998 Wis. L. Rev. 131, 132–33.

27. See generally Patrick Atiyah, The Damages Lottery 143–50 (1997) (describing the range of damage awards as a "lottery"); Michelle Chernikoff Anderson & Robert J. MacCoun, Goal Conflict in Juror Assessments of Compensatory and Punitive Damages, 23 Law & Hum. Behav. 313, 327–28 (1999) (discussing two studies that suggested jurors often fail to "compartmentalize" compensatory and punitive damages, leading to "leakage" between the two categories).

28. See, e.g., Jonathan M. Karpoff & John R. Lott, Jr., On the Determinants and Importance of Punitive Damages Awards, 42 J.L. & Econ. 527, 540–45 (1999) (suggesting that punitive damage awards vary in ways not explained by injury or defendant characteristics); Leebron, supra note 13, at 309–11 (discussing the implications of variability in awards).

"noise," unexplained by objective factors.[29] A study of all reported sexual harassment cases was unable to connect either compensatory or punitive awards to any case characteristics that might be thought to explain jury judgments.[30] The punitive damage area is more complicated—a point to which we will return shortly—but there is evidence of significant variability here as well.[31] The most ambitious claims to the contrary attempt to show that once the compensatory award has been made, the punitive award becomes predictable to a certain degree;[32] but the same data show that at the time a case is filed (before the amount of compensatory damages is known), it is very hard to know the expected punitive award, and that there is generally a great deal of "noise" in outcomes.[33]

To understand the current study, it is necessary to understand its predecessor by way of background. Our earlier study involved a demographically diverse set of jury-eligible citizens from Travis County, Texas.[34] The relevant experiment involved twenty-eight personal injury cases, which respondents were asked to assess in one of three ways: outrageousness, on a rating scale (0 to 6); intent to punish, on a rating scale (also 0 to 6); and actual awards, on the unbounded scale of dollars. As noted, our principal findings were twofold. People's moral judgments are widely shared and predictable—in fact strikingly so—at least in the personal injury cases investigated in this study.[35] But in spite of this point, and in the presence of shared moral judgments, people's judgments on a dollar scale—the scale, or "response mode," favored by the legal system—are highly unpredictable, in the sense that the median

29. See Leebron, supra note 13, at 310.

30. See Shih & Sunstein, supra note 12 (manuscript at 1–3).

31. See Karpoff & Lott, supra note 28, at 540–45. There is some dispute over the degree of unpredictability. Eisenberg et al., supra note 10, show that the logarithm of punitive awards is predicted reasonably well from a set of objective characteristics of cases in which awards were made; in particular, the authors show that the compensatory award is a fairly good predictor of the punitive award. See id. at 644. But the authors themselves note that the range of possible awards in regular (i.e., not logarithmic) dollars is still quite high even after controlling for the many factors in their regression model, including compensatory damages. For example, in their data, for a case with a $500,000 compensatory award, 5% of punitive damage awards would be $10,000 or less, but another 5% would be $6,500,000 or more. See id. at 657; see also Karpoff & Lott, supra note 28, at 540–45 (discussing the difficulty of predicting punitive damage awards and the consequences of that difficulty).

Note also that predictability can be understood in different ways: (a) predictability exists when case characteristics predict punitive awards; (b) predictability exists when the judgments of one group of six or twelve predicts the judgments of another group of six or twelve; (c) predictability exists when an actor can assess expected liability when something goes wrong. Our principal emphasis here is on (b); Eisenberg's emphasis is on (a); both are relevant to (c). Of course the three are closely related in practice. We offer a more detailed treatment of predictability below. See infra Part II.B.8–9.

32. See Eisenberg et al., supra note 10, at 637–39.

33. See Karpoff & Lott, supra note 28, at 543.

34. See Sunstein, Kahneman, & Schkade, supra note 4, at 2094–108.

35. See id. at 2097–99.

judgment of any group of twelve people is an extremely poor predictor of the median judgment of any other group of twelve people.[36] Lacking a reliable understanding of how deliberation would affect individual judgments, we used the median of groups of twelve individuals, randomly selected from our pool of 899 citizens and combined them into a large number of statistical juries.[37] At least in this setting, the primary identifiable source of the noise is the difficulty jurors have in translating their punitive intent into dollars. Dollar awards are highly variable despite the existence of shared moral judgments.

To explain why the use of the dollar scale would produce variability, we developed a theory of juror punitive damage judgments. The *outrage model* assumes that a juror's basic response to a defendant's behavior is a reaction of outrage, which in turn leads to an intent to punish, which can be expressed on different response scales (for example, a dollar amount or a rating from 0 to 6). These scales vary not only in their complexity, but also in the precision and consistency of the measurements that they provide: Some scales are less reliable than others, in the sense that they are less consistent at producing the same answer to the same question, or different answers to different questions. As we have already seen, the dollar scale is in this sense an extremely unreliable expression of punitive intent, and it produces a high degree of arbitrariness.

To understand the reasons for the noise in dollar damage judgments, we explored the close analogy between our findings with respect to the dollar scale and the outcome of psychological research on the problem of "magnitude scaling," which occurs when people are asked to indicate the intensity of their subjective responses to stimuli—the brightness of lights, the loudness of noise—along an unbounded numerical scale.[38] In some of these experiments, the participants are given a "modulus," which specifies the number that is to be assigned to a particular standard stimulus. In other experiments, the participants are not given a modulus. In the absence of a modulus, variability increases dramatically; some participants assign high numbers, others assign low numbers. With the dollar scale, the underlying problem is that people are being asked to scale without a "modulus," that is, without a standard that would help give meaning to various numbers on the scale.[39]

---

36. See id. at 2100–03.

37. We relied on evidence suggesting that the median judgment of a group of predeliberative individuals is a good predictor of the judgment that group will reach as a result of deliberation. See id. at 2101 nn.127–28, citing James H. Davis, Group Decision Making and Quantitative Judgments: A Consensus Model, *in* 1 Understanding Group Behavior 35, 47 (Erich H. Witte & James H. Davis eds., 1996); Shari Diamond & Jonathan Casper, Blindfolding the Jury to Verdict Consequences, 26 L. & Soc'y Rev. 513, 553 (1992). We noted, however, the possibility of effects of the sort we observed in the current study. See Sunstein, Kahneman, & Schkade, supra note 4, at 2101 n.128 (noting that deliberation may result in more extreme awards due to "amplification of bias").

38. See S.S. Stevens, Psychophysics 25–31 (1975).

39. See Sunstein, Kahneman, & Schkade, supra note 4, at 2106–07.

The key point is that when a modulus is supplied, the variability greatly decreases; in its absence, respondents adopt their own moduli. Juries asked to assess punitive damage awards are in effect asked to scale without a modulus. Unpredictable judgments are a natural result even when people do not disagree about the significant issues[40]—even when, that is, there is a kind of "bedrock" moral judgment in place.[41] If this point is correct, it helps explain the observed variability in dollar awards in many areas of the law. It also helps explain the disparities that led to the enactment of the Sentencing Guidelines;[42] before the guidelines, judges were being asked, in effect, to scale without a modulus, since the relevant scale (years) has a great deal in common with the dollar scale (i.e., bounded below at zero, but with great discretion at the high end).

Our earlier study did not, however, involve deliberating juries, and a natural question was whether deliberating juries would produce similar or quite different results. Perhaps deliberation would reduce variability.[43] An alternative possibility was that the process of collective delibera-

---

40. For similar results in the context of compensatory pain and suffering awards, see Michael J. Saks et al., Reducing Variability in Civil Jury Awards, 21 Law & Hum. Behav. 243, 243–44 (1997).

41. A similar point is made in Lane v. Hughes Aircraft Co., 93 Cal. Rptr. 2d 60, 72 (2000) (Brown, J., concurring in the result) (citation omitted):

> Significantly, the variability in punitive damage awards does not flow so much from any inherent inability of different juries to agree on the wrongfulness of specific conduct. Rather, it results from the way courts ask juries to measure that wrongfulness. . . . The variability in punitive damage awards does not, therefore, reflect any inconsistency in jury factfinding, merely that awards are not calibrated to a common scale.

42. See Sandra Shane-DuBow et al., U.S. Dep't of Justice, Sentencing Reform in the United States 7 (1985).

43. In a series of papers, Michael Saks has argued that juries actually reduce variability. Saks's early research compared twelve-person juries to six-person juries, see the overview in Michael J. Saks, Jury Verdicts, 61, 77–92 (1977)—a comparison on which our study here does not bear. Saks subsequently extended his research to include a comparison between juries and judges, with the suggestion that juries are likely to produce less variability by virtue of their numbers. See Michael J. Saks, Do We Really Know Anything About the Behavior of the Tort Litigation System—and Why Not?, 140 U. Pa. L. Rev. 1147, 1269, 1271–74 (1992) [hereinafter Saks, Tort Litigation System]; 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, 850 & n.204 (1992); Saks et al., supra note 40, at 243–44, 246. This conclusion is briefly challenged in Robert MacCoun, Inside the Black Box: What Empirical Research Tells Us About Decisionmaking by Civil Juries, *in* Verdict: Assessing the Civil Jury System 137, 178 n.126 (Robert E. Litan ed., 1993) ("The argument is based on statistical sampling theory, but the analogy between empaneled juries and random samples is an imperfect one. Though it is a plausible hypothesis, it requires more rigorous testing than it has received to date."). We have attempted a more rigorous test here, finding that juries produce more variability as compared with the mean of individual predeliberation judgments. This finding suggests, though it certainly does not prove, the possibility that juries will produce more variability in awards than judges (a suggestion supported by the possibility that judicial experience with a wide range of cases will introduce the equivalent of a "modulus" by which to discipline dollar awards).

tion would move the group further in the direction of the initial tendency suggested by the individual judgments. In any case, a test of deliberating juries would help to confirm or deny the wisdom of the decision to treat the median judgment of a group of twelve as the likely judgment of any deliberating group (for purposes of creating statistical juries). Hence our main purpose in this Empirical Study was to examine the effects of jury deliberation on dollar awards and, in particular, to see whether deliberation would increase or decrease predictability. In the process, we also hoped, as a secondary goal, to find out whether the original findings—shared moral judgments but erratic awards—would be replicated with a new sample of citizens from a different state, and with new and richer case materials.

II. Deliberating Juries: An Experimental Inquiry

A. *Method*

Jury-eligible citizens from Phoenix, Arizona were recruited by a survey firm and paid $35 for their participation. Each juror was randomly assigned to a six-person jury, and each jury was assigned to a response mode order; half of the juries judged dollar awards first and punishment ratings second, and the other half completed the tasks in the opposite order. Each jury judged only one case, which was the subject of both its punishment rating (on a scale of 0 to 8) and its dollar award. Six juries (out of a total of 480) had only five members because an insufficient number of participants showed up at a given appointment time. A pilot test of twenty-nine juries was conducted in Phoenix to test the materials and procedure. Because adjustments were very minor, these juries were added to the main sample and the combined sample was analyzed together. Therefore, a total of 3048 citizens participated in 509 juries.

The procedure consisted of four parts. In Part 1, all participants in a given session viewed a videotape for the case they would consider, read the corresponding written materials, and recorded their personal judgment of the appropriate punitive damage award or punishment rating. In Part 2, participants were randomly assigned to juries of six members, which were given thirty minutes to deliberate on and reach a unanimous verdict on a punitive damage amount or a punishment rating. In Part 3, a new individual response form was distributed, which asked participants to record a second personal judgment for the same case, using the type of verdict (punishment rating or dollar damages) complementary to the one they had already used. In Part 4, each jury again deliberated to reach a unanimous verdict on this second type of judgment for the same case. Thus, for each individual, and for each jury, we have *both* a dollar award *and* a punishment rating for the case they considered. We use the terms dollar and punishment *judgments* to refer to the dollar awards and punishment ratings made by individuals. For juries we will refer to these as dollar and punishment *verdicts*. For purposes of understanding real-world behavior, the dollar awards are most important. We inquire into

punishment ratings both to understand the relation between punishment ratings and dollar awards, and to see the effect of deliberation on both of these.

TABLE 1. RESPONSE MODE MANIPULATION

Punishment

How much should the defendant be punished because of their actions, and to deter the defendant and others from similar actions in the future? Note that the compensatory damages that the defendant must pay do not count as part of the punishment. Please circle the number that best expresses the *jury's* judgment of the *appropriate level of punishment.*

None		Mild		Substantial		Severe		Extremely Severe
0	1	2	3	4	5	6	7	8

$ Damages

What amount of punitive damages (if any) should the defendant be required to pay as punishment and to deter the defendant and others from similar actions in the future? Note that the compensatory damages that the defendant must pay do not count as part of the punishment. Please write the *amount of punitive damages* that the *jury* agreed on in the blank below.

$ _____

The case materials consisted of fifteen personal injury scenarios (summarized in Table 2).[44] An example is provided in the Appendix. A videotape was prepared for each case, in which a professional actor read the text of the case and all instructions aloud. To maximize comprehension, participants were required both to view the videotape and to read the written version. The size of the defendant firm (annual profits of $100–$200 million) and compensatory damages ($200,000) were the same for all cases. Thus, the variability we observe cannot be accounted for by a model that depends on variability in compensatory damage awards or in the defendant's ability to pay.

B. *Results*

1. *Preliminaries.* — Notwithstanding the half-hour time limit for deliberation, 91% of juries reached a unanimous verdict on a punishment rating (a total of 461 verdicts) and 82% of juries reached a unanimous

44. Of these, 10 were more elaborate versions of the same scenarios used in Kahneman, Schkade, & Sunstein, supra note 4, at 79–84, and five were completely new scenarios which, like the first 10, were based on real cases (Table 2). The main substantive elaboration on the original scenarios was the addition of a paragraph of closing arguments by the attorneys for each side.

TABLE 2. SUMMARY OF PERSONAL INJURY SCENARIOS

Case	Description
Williams v. National Motors	Motorcycle driver injured when brakes fail
Smith v. Public Entertainment	Circus patron shot in arm by drunk security guard
Douglas v. Coastal Industries	Auto airbag opens unexpectedly, injuring driver
Sanders v. A&G Cosmetics	Man suffers skin damage from using baldness cure
Stanley v. Gersten Productions	Elderly woman suffers back injuries from using exercise video
Glover v. General Assistance	Child ingests large quantity of allergy medicine, needs hospital stay
Lawson v. TGI International	Employee suffers anemia due to benzene exposure on the job
Newton v. Novel Clothing	Small child playing with matches burned when pajamas catch fire
West v. MedTech	Disabled man injured when wheelchair lift malfunctions
Windsor v. Int'l Computers	Secretary chronically ill due to radiation from computer monitor
Reynolds v. Marine Sulfur	Seaman injured when molten sulfur container fails
Crandall v. C&S Railroad	Train hits car at crossing, injuring driver
Dulworth v. Global Elevator	Shopper injured in fall when escalator suddenly stops
Hughes v. Jardel	Store employee raped in mall parking lot
Nelson v. Trojan Yachts	Man nearly drowns when defective boat sinks

verdict on a dollar award (a total of 416 verdicts). The remainder had not reached a verdict when the time limit expired; these were treated as hung. All further analyses were conducted on the 401 juries that reached *both* a punishment verdict *and* a dollar verdict.[45] Because there were no statistically significant differences between the verdicts of juries that judged dollars first and those that judged punishment first, we analyzed together the verdicts made by dollar-first juries and punishment-first juries.

2. *Overview: How Do the Verdicts of Deliberating Juries Compare to Those of Statistical Juries?* — We assessed the effect of deliberation on juror judgments by comparing each jury's verdict to the median predeliberation

45. We chose the more conservative path of focusing on juries with complete responses to ensure that comparisons between punishment and dollar verdicts, and between individuals and juries, were based on the same set of respondents. Recreating our Tables and Figures with all available responses produces the same pattern of results, with some slight differences in exact numbers.

TABLE 3. MEDIAN VERDICTS FOR DELIBERATING AND STATISTICAL JURIES

	Punishment Verdicts			Dollar Verdicts		
Case	Statistical Juries	Deliberating Juries	Average DSM	Statistical Juries	Deliberating Juries	Average DSM
Reynolds	5.5	6.0	15%	1,875,000	10,000,000	54%
Glover	5.0	5.0	1%	1,000,000	4,000,000	52%
Lawson	4.3	4.5	4%	475,000	2,000,000	53%
Williams	5.0	5.0	14%	550,000	1,500,000	46%
Smith	5.5	6.0	19%	325,000	1,000,000	52%
Nelson	5.0	5.0	20%	450,000	1,000,000	48%
Hughes	5.0	5.0	12%	450,000	1,000,000	45%
West	4.5	5.0	9%	500,000	1,000,000	34%
Douglas	4.0	4.0	11%	225,000	500,000	40%
Crandall	4.0	4.0	-8%	200,000	500,000	35%
Sanders	3.5	3.0	-8%	50,500	100,000	25%
Windsor	3.0	2.0	-26%	37,500	50,000	38%
Stanley	1.0	1.5	0%	0	0	0%
Dulworth	0.3	0.0	-15%	0	0	17%
Newton	0.0	0.0	3%	0	0	23%
Mean of Top 5	5.1	5.3	11%	845,000	3,700,000	51%
Mean of Middle 5	4.5	4.6	9%	365,000	800,000	40%
Mean of Bottom 5	1.6	1.3	-9%	17,600	30,000	21%
Overall Mean	3.7	3.7	3%	409,200	1,510,000	37%

judgment of the individuals who composed that jury. We will refer to the median predeliberation judgment of the individuals in a jury as the verdict of the *statistical jury*. To evaluate the effects of deliberation, we compare the verdicts of *deliberating juries* with those of statistical juries.

The results observed for the fifteen cases are shown in Table 3.[46] The cases are arranged in the Table in descending order of the median dollar verdict of deliberating juries. Note first that the median verdicts of deliberating and statistical juries produce very similar *rankings* of the cases. For dollars, there is a Spearman rank correlation[47] of .88 between the deliberating and statistical jury verdicts in Table 3; for punishment verdicts the average rank correlation is even higher, at .98. The correlation between punishment verdicts and dollar verdicts is also high, at .87. These results confirm the finding of earlier research that, in the aggregate, judgments of punitive intent and of dollar awards share the same core of moral outrage, and therefore produce the same ordering of cases.

While there is agreement on the ordering of cases, the level of verdicts tells a different tale. Punishment verdicts are, on average, quite close for statistical and deliberating juries, but dollar verdicts show a dra-

46. The columns labeled DSM are explained below, at Part II.B.4.

47. The Spearman rank correlation is an index of agreement between rankings that is analogous to first converting each column to ranks (from I to 15 in this case) and then computing the correlation between the two sets of ranks.

matic difference: Deliberating juries produce much higher awards, especially but not only at the high end. Indeed, 83% of the 330 non-zero dollar verdicts were above the median individual judgment for that jury. This is the most important finding in the study: the severity shift in dollar verdicts.

In summary, then, aggregate verdicts from deliberating and statistical juries show strong agreement on the relative egregiousness of the cases, and for punishment verdicts, they do not dramatically diverge. Deliberating juries, however, produce dollar verdicts that far exceed the median judgments of the jurors that compose them. We now try to understand how this pattern might occur. To do so, we divide verdicts into three decisions: (1) the decision about whether to punish at all; (2) the decision about the appropriate punishment verdict; and (3) the decision about the appropriate dollar verdict. As we shall see, the effects of deliberation are quite different for each decision.

3. *Punish or Not Punish: A Majority Model.* — The first decision for a jury is, presumably, whether to punish or to reject punishment by a verdict of $0 in damages or a 0 punishment rating. Table 4 shows the percentage of non-zero verdicts that were made by juries, in relation to the initial distribution of judgments among the jurors. The pattern is identical for punishment and dollar verdicts: When a majority of juror judgments (i.e., four or more) are 0, the jury verdict is virtually certain to itself be 0. When a majority of jurors have non-zero judgments, the jury verdict is virtually certain not to be 0. Finally, if the jury is evenly split, the chance of a 0 verdict is about 50–50.

Without detailed analysis of the deliberation transcripts, we do not know whether juries actually voted or explicitly agreed to adopt a majority decision rule. We observe only that the pattern of results is consistent with the adoption of such a rule. In contrast to other phases of the jury decision that we consider later, there is no evidence of any systematic effect of deliberation on outcomes (i.e., juries were neither more nor less likely to punish than their jurors). Thus, for the decision of whether or not to impose punitive damages, there is no indication of any asymmetry of power or influence between jurors who were initially inclined to say yes and those who were inclined to say no.

TABLE 4. PERCENTAGE OF NON-ZERO VERDICTS AS A FUNCTION OF PREDELIBERATION JUDGMENTS

	Jury Verdicts	
Individual Predeliberation Judgments	Non-Zero Punishment Ratings	Non-Zero $ Awards
Majority non-zero	99%	98%
Even split	48%	45%
Majority zero	8%	4%

4. *Deliberation Shift Analysis.* — We now turn to the severity of punishment verdicts chosen by the juries that determined that some punishment was appropriate. We wish to examine the relationship between the postdeliberation verdict of a jury and the predeliberation distribution of judgments among its members. For this purpose we introduce a *deliberation shift analysis,* which we will apply to both punishment ratings and dollar awards. The predeliberation judgments of jurors are first ranked, from the most lenient to the most severe; the eventual verdict of the jury is then inserted in that ordering, and its rank is computed. For example, suppose that the individual jurors had predeliberation judgments of $0, $200,000, $300,000, $500,000, $1,000,000, and $5,000,000, and that the jury verdict was $750,000. The jury verdict ranks fifth in the distribution of individual judgments of its members. In this instance, the jury was more severe than four of its original members, and less severe than two of its members, indicating that, overall, deliberation made judgments more severe.

If the outcomes of deliberation were determined by a simple voting model, the jury verdict would always be in the middle of the distribution of initial judgments, at the median. There would be no shift, either toward greater leniency or toward more punishment. With no shift, for a jury of six (with the jury verdict added as the seventh member), the predicted position of the jury in the distribution of the opinions of its members is always fourth.[48] The *deliberation shift measure (DSM)* is the *difference* between the observed and the predicted rank of the jury verdict, as a percentage of the maximum possible shift in the direction taken. To continue our dollar example above, since the jury verdict ranks fifth among its jurors' predeliberation judgments, the difference would be $5 - 4 = 1$. For a jury of six, the maximum possible upward shift is $7 - 4 = 3$, and the DSM would be 1/3, or 33%. This means that the rank of the jury verdict was 33% of the way from the rank of the median juror (4) to the rank of the maximum juror (7).[49] The DSM is positive if the jury is more severe than its median member; it is negative if the jury is more lenient than its median member. If the jury verdict was higher than the judgment of the maximum juror, the DSM would be 100%; if it was lower than the judgment of the minimum juror, the DSM would be -100%. To study the systematic effects of deliberation, we computed the DSM for every nonhung jury, separately for punishment verdicts and for dollar verdicts. Table 3 shows the mean values of the DSM for each of the fifteen cases, for both punishment and dollar verdicts.

5. *Punishment Ratings Either Up or Down.* — For punishment verdicts there is a clear pattern in the results, which can be observed both in the column of DSM values and by comparing the statistical and deliberated

48. For a jury of 12, the expected rank would be 7; for a jury of nine, the expected rank would be 5.5; and so forth.

49. Because the DSM is formulated as a percentage, it can be computed for, and has the same interpretation for, a jury of any size.

verdicts: Deliberation increased the severity of punishment for high-punishment cases and reduced it for low-punishment cases. Reading down the table, the DSM is positive for nine of the top ten cases, and negative for four of the bottom five cases. There was a severity shift for the high-punishment cases, and a leniency shift for the low-punishment cases.

Because the table is arranged roughly in decreasing order of punitive intent, we can see that the DSM is positive for high-punishment cases (average for the top ten cases is 10%) and negative for low-punishment cases (average for the bottom five cases is -9%). Further, the correlation between the DSM and the median statistical jury verdicts is .67, which means that the more severe the individual predeliberation judgments, the greater the shift. In the language of the group polarization literature, we observe systematic *choice shifts*, in which deliberation generally increases differences among cases, by making severe verdicts more severe and lenient verdicts more lenient, relative to the predeliberation judgments of jurors.

6. *Dollar Awards and the Severity Shift: Deliberation Increases Punitive Damages.* — We now turn to the task of understanding the remarkable difference between the dollar awards obtained from deliberating juries and those obtained from a statistical pooling of the predeliberation opinions of jury members. The basic result is that deliberation causes awards to increase, and it causes high awards to increase a great deal. As extreme but actual illustrations of the severity shift, consider a few examples from the raw data:

- A jury whose predeliberation judgments were $200,000, $300,000, $2 million, $10 million, $10 million, and $10 million reached a verdict of $15 million.
- A jury whose predeliberation judgments were $200,000, $500,000, $2 million, $5 million, and $10 million reached a verdict of $50 million.
- A jury whose predeliberation judgments were $2 million, $2 million, $2.5 million, $50 million, and $100 million reached a verdict of $100 million.

Now consider the DSM column for dollar verdicts in Table 3. Recall that the value of the DSM is positive if the jury verdict is more severe than the median judgment of its jurors and negative if the jury is more lenient. The pattern of results is clear: The DSM is generally positive, indicating that deliberation generally produced a severity shift. Furthermore, the DSM for dollar verdicts is much higher for high-punishment cases than for low-punishment cases: The correlation between the median punishment verdict and the DSM for dollar verdicts is .95.

The difference between deliberating and statistical juries is very large, especially for the high-punishment cases: For the top ten cases in Table 3, the average DSM of 46% means that the jury verdict is about halfway between the second-highest and third-highest individual judgments. Even more surprising, for the ten high-punishment cases, 10% of jury verdicts were *even higher* than the highest individual judgment (i.e.,

the DSM was 100% for these juries). A further 17% of verdicts were equal to the highest individual judgment (i.e., a DSM of 83%). These extreme verdicts were less common for the five low-punishment cases, in which 15% of verdicts equaled the highest individual judgment, and none exceeded this maximum. The pattern is clear: Deliberation made dollar verdicts more severe, especially for high-punishment cases.

Notably, we did not find that the degree of dispersion between individual predeliberation judgments contributed to greater or lesser shifts as a result of deliberation. For example, for juries with non-zero verdicts for the same case, the average correlation between the standard deviation of individual judgments (a measure of dispersion) and the DSM was –.05 for dollars and .08 for punishment (neither correlation is statistically different from 0). In other words, juries whose members were in rough agreement (i.e., had a low standard deviation) about dollars or punishment did not show a different shift from groups whose members were in substantial disagreement about dollars or punishment.

7. *Do People from Arizona Agree with People from Texas? The Effects of Geography, Race, Gender, Education, Age, and Wealth.* — A subsidiary but nonetheless important question is whether the findings of the earlier study were replicated under the current study's changes in stimuli, procedure, and sample. The answer is that the previous results were replicated in every essential respect. The findings in the Texas study were replicated in Arizona, and despite evident differences between the two regions, people from the two areas evaluated cases in the same way. As before, dollars and ratings produced very similar rankings of the cases (a rank correlation of .90 compared to .91 for the comparable condition in the previous study[50]). Different demographic groups again produced very similar average evaluations, as indicated by the extremely high correlations in Table 5.

In addition, the ordering of case evaluations closely matched that in our previous study. There are ten cases common to both studies, and evaluations made by Texans in the previous study were highly predictive of those made by Arizonans in the current study—the rank correlation between the two samples was .90 for punishment ratings and .98 for dollar awards. Thus, the current, larger study, with several nontrivial changes, confirmed the conclusion of our previous study that individual moral judgments are predictable and shared, but expressing them in dollars produces unpredictability and confusion.

50. This correlation was computed for the condition in the previous study, see Sunstein, Kahneman, & Schkade, supra note 4, that is directly comparable to the current study, which contained cases with large companies and high harm.

TABLE 5. CORRELATION BETWEEN DEMOGRAPHIC GROUPS ON INTENDED SEVERITY OF PUNISHMENT*

		Men		
Gender	*Women*	.99		
		White	*Hispanic*	
Ethnicity	*Hispanic*	.92		
	Other	.88	.81	
		<$30K	*$30–50K*	
Household Income	*$30–50K*	.98		
	>$50K	.99	.99	
		<30	*30–39*	*40–49*
Age	*30–39*	.97		
	40–49	.96	.97	
	>50	.96	.97	.97

* Entries are correlations between mean responses to scenarios by respondents in the indicated demographic categories.

8. *With Respect to Dollars, How Predictable Are Jury Verdicts?* — An important goal of the legal system is to treat the similarly situated similarly. Our previous study showed that both the dollar judgments of individuals and the dollar verdicts of statistical juries would probably fail this test of procedural justice, because of a high degree of unpredictability in damage awards for the same case, as well as inconsistency in distinguishing between cases of more and less egregious conduct.[51] Among many in the legal community there is the hope, and indeed the conviction, that deliberation by a group of jurors will overcome individual biases and produce more just and more predictable verdicts. As will be seen, our findings lend no support to this view.

The simplest and most practical criterion for predictability is reflected in the distribution of possible verdicts for a given case (a criterion that asks the extent to which the *identically* situated are treated identically). This is of course a critical piece of information for a lawyer advising a client about whether or not to settle a dispute, or for an actor contemplating liability for a potentially tortious course of conduct. In our sample, we had multiple independent juries rendering punitive damage verdicts for the same case, and this information can be used to estimate verdict distributions. Table 6 presents selected distributional statistics for each case. The range of possible dollar verdicts is strikingly large. For example, each of the top five cases has a minimum award of $500,000 or less, and yet the average maximum award for these cases is over

51. See id. at 2077–78, 2100–03.

$83,000,000. Further, the maximum verdicts are 10 to 500 times as large as the median verdicts (for cases with non-zero medians). Even for the three cases at the bottom with zero medians (i.e., a majority of juries for that case awarded no punitive damages), plaintiffs could still be awarded $500,000. Also, although there is considerable noise (in part because the number of juries for each case is relatively small), the range of verdicts for a given case tends to increase in proportion to the median verdict.[52] Note that these variations between juries occurred on identical presentations of identical facts, unaffected by differences in (for example) compensatory awards or lawyers' presentations.

To make the uncertainty of these dollar verdicts more concrete, imagine that a statistically sophisticated and greatly experienced lawyer is advising a defendant about a possible punitive damages award, on the basis of the data illustrated by Table 6. For the purpose of the illustration, assume that the lawyer is not only sophisticated but also wise, and able to make unbiased predictions of jury decisions: When she states that her best guess is an award of $X, the actual award is equally likely to be above or below X (this is her estimate of the median award). On the basis of our data the lawyer would be able to provide the client with the following information:

> "My best guess is that you will face a judgment of $X. There are equal chances that it will be higher or lower than this amount. However, there is a lot of uncertainty about how much higher or lower it will be: There is a 10% chance that you will have to pay more than a times that amount, and there is a 10% chance that you will have to pay less than $1/b$ of that amount."

Averaging across cases, the best estimates[53] of a and b, respectively, are 7.74 and 6.61. On the basis of these values, a lawyer who predicts a verdict of $2 million should also estimate that there is a 10% chance that the actual verdict will be over $15.48 million, and a 10% chance that it will be less than $0.30 million. Because the range increases proportionately with the median (except for noise), the same values of a and b apply for any value of $X.[54] Finally, these estimates assume a jury of six. The uncertainty would very likely be reduced somewhat with a larger jury.[55]

---

52. We observed a similar pattern in our previous study. See Kahneman, Schkade, & Sunstein, supra note 4, at 69.

53. To obtain these estimates, we computed the 90th percentile/median and median/10th percentile ratios for each case, and then computed the geometric mean across cases for each ratio. The estimates reported here are for the nine cases that have neither a median of zero nor a 10th percentile of zero (see Table 6).

54. To test for proportionality, we ran a regression of the difference between the 90th and 10th percentiles for a given case on the median for that case. If the range goes up proportionately with the median, then this regression should have a good fit, and the constant in the regression should be close to zero. In fact, the line fits quite well ($R^2 = .66$) and the constant is not significantly different from zero ($p > .05$).

55. Because statistical uncertainty is proportional to the size of the jury, we can approximate how much smaller a and b would be for a jury of 12, under the assumption that uncertainty in deliberating juries would diminish at the same rate as in a statistical

TABLE 6. PERCENTILES OF JURY DOLLAR VERDICTS, BY CASE (IN THOUSANDS OF DOLLARS)

	Minimum			Median			Maximum
Case	0th	10th	25th	50th	75th	90th	100th
Reynolds	250	1000	3500	10,000	17,500	50,000	100,000
Glover	500	1000	1250	4000	10,000	50,000	100,000
Lawson	200	250	1000	2000	6000	15,000	100,000
Williams	100	200	700	1500	5000	10,000	15,500
Smith	0	100	300	1000	7000	20,000	100,000
Nelson	100	250	500	1000	5000	5000	100,000
Hughes	0	200	850	1000	2000	20,000	40,000
West	1	200	500	1000	2000	4000	10,000
Crandall	0	50	250	500	1450	2000	100,000
Douglas	0	1	250	500	1000	25,000	50,000
Sanders	0	0	0	100	500	1000	50,000
Windsor	0	0	0	50	400	5000	25,000
Newton	0	0	0	0	75	200	500
Dulworth	0	0	0	0	40	300	500
Stanley	0	0	0	0	25	250	500
Mean of Top 5	210	510	1350	3700	9100	29,000	83,100
Mean of Middle 5	20	140	470	800	2290	11,200	60,000
Mean of Bottom 5	0	0	0	30	208	1350	15,300
Overall Mean	77	217	607	1510	3866	13,850	52,800

9. *Are Deliberating Juries More Predictable Than Statistical Juries?* — In the Introduction, we asked whether deliberating juries would produce dollar verdicts that are more predictable than those of statistical juries. We can now use the *a* and *b* analysis set forth above[56] to answer this question. As with the verdicts of deliberating juries, variability in the verdicts of statistical juries is roughly proportional to the median verdict.[57] We can apply the same procedure as before to estimate the factors *a* and *b*, which measure the estimated relationship of the 10th and 90th percentiles for each case to the median award for that case.

For statistical juries, our estimates of *a* and *b* are 2.88 and 4.11, which are both far lower than the corresponding figures of 6.61 and 7.74 for deliberating juries. In our example above, the lawyer's predicted range for a statistical jury verdict would be from $0.69 to $8.22 million compared to the range for a deliberating jury verdict of $0.30 to $15.48 mil-

jury. In this case, since the jury would be twice as large, we divide *a* and *b* by the square root of 2. The resulting estimates are $a = 4.67$ and $b = 5.47$, which in the lawyer advice example would produce a predicted verdict range of $0.43 million to $10.94 million.

56. See supra Part II.B.8.

57. To test for proportionality, we ran a regression of the difference between the 90th and 10th percentiles of statistical jury verdicts for a given case on the median for that case, just as we did for deliberative jury verdicts, as discussed supra at note 54. Again, the line fits fairly well ($R^2 = .56$), and the constant is not significantly different from zero ($p > .05$).

lion. Obviously, there is far less uncertainty about the verdicts of statistical juries than about those of deliberating juries.

This pattern is remarkably consistent across cases. There is greater uncertainty in deliberating jury verdicts than in statistical jury verdicts for each of the ten cases for which *a* can be calculated (those with non-zero medians), and greater uncertainty for eleven of the twelve cases for which *b* can be calculated (those with non-zero 10th percentiles). It is important to note that this estimation procedure effectively controls for the severity shift, and therefore that these differences are not due merely to the generally higher level of verdicts by deliberating juries. We conclude, rather to our surprise, that deliberation is a significantly poorer way of aggregating opinions than is statistical pooling—at least if the goal is to decrease the arbitrary unpredictability of awards.

III. WHAT HAPPENED? SEVERITY SHIFTS, RHETORICAL ASYMMETRY, AND RELATED PHENOMENA

We now turn to a discussion of these results. We emphasize three phenomena. The first is identified for the first time here, while the second and third have been studied in many previous experiments.

- The first and most important phenomenon is the severity shift. We believe that this occurred because of a *rhetorical asymmetry* that gives one set of arguments an automatic, other-things-equal upper hand in a group discussion, so that groups will typically shift in the direction holding that upper hand.
- The second phenomenon, described standardly though somewhat vaguely as a "choice shift," occurs when the decision of a group shifts toward a more extreme version of the view held, before deliberation, by the group's median member.[58] Our evidence clearly shows choice shifts with respect to punishment verdicts, and because high awards increased much more than low awards, we think that something similar played a role in dollar verdicts as well.
- The third phenomenon, known as "group polarization," occurs when individuals, polled privately after group discussion, shift toward a more extreme point in the direction set by the original distribution of views.[59] Because our jurors were not polled privately after discussion, we do not have direct evidence of group polarization, though there is reason to believe that it may have occurred.[60]

---

58. See, e.g., Zuber et al., supra note 3, at 50.

59. See id. Choice shift and group polarization ordinarily accompany one another, although it is possible to have one kind of movement without the other. See id. at 59.

60. See Roger Brown, Social Psychology 229 (2d ed. 1986) ("In every [mock jury study] where the report of data makes it possible to check, group polarization occurs." (citations omitted)). Brown infers group polarization from the fact that "for ninety percent of juries that must reach unanimous agreement and do not hang, the final verdict is consistent in direction with the majority on the initial ballot." Id. Similarly, it is

A. *The Severity Shift and Rhetorical Asymmetry*

1. *Rhetorical Asymmetry.* — By far the most striking finding in our data is the severity shift produced by deliberation. What mechanism causes a jury to decide on an award that exceeds the initial judgment of its median member—and sometimes to exceed the highest predeliberation judgment of all its members?

We hypothesize that a feature of deliberation, a rhetorical asymmetry, helps produce the one-way movement that we observe. Specifically, we hypothesize that once the jury has agreed that there will be a non-zero dollar award, the arguments for a larger award have a rhetorical advantage and are more persuasive. If this is the case, then a jury would be drawn disproportionately toward the larger predeliberation judgments of its jurors. No such asymmetry would be expected for the punishment scale, if it is hypothesized that social norms give the advantage, not to anyone arguing that the conduct of a corporate defendant was "worse" in the abstract, but to anyone arguing for a higher dollar award against a corporate defendant. The key point has to do with the translation of a punishment judgment into a dollar award; those who argue that "more" money is necessary to punish a corporation appear to have the upper hand. The unbounded dollar scale affords great latitude in the expression of what "more" means.

To examine the hypothesis of rhetorical asymmetry more directly, we asked eighty-seven University of Chicago law students whether it would be harder to argue for a smaller or a larger award. In this study, respondents were simply told that they were deliberating about punitive damage awards and were given no details of any particular case. They were first asked to generate arguments for a higher or lower award, and then asked which award (higher or lower) would be easier to justify. Half of the students were asked to argue for a higher award; half were asked to argue for a lower one. After generating the relevant arguments, they were asked to complete a second task, presented as follows:

> Imagine that a jury in a civil trial is deliberating about a personal injury case in which the defendant is a large corporation (with annual profits of approximately $200 million). The jury has already (a) unanimously ordered the defendant to pay an amount of compensatory damages that fully compensates the plaintiff, and (b) unanimously concluded that while the underlying conduct was not truly horrendous, it was sufficiently reckless to justify an award of punitive damages as well (in addition to compensatory damages).

---

reasonable to infer, from the dramatic choice shifts we observe, that group polarization is highly likely, in the sense that individuals will likely have shifted in the direction indicated by the group's decisions.

For a general overview and discussion, with many applications to legal problems, see Cass R. Sunstein, Deliberative Trouble? When Groups Go to Extremes, 110 Yale L.J. (forthcoming Oct. 2000).

> In general, which position would you expect to be harder for a juror to argue for in a deliberation? (please circle the letter of your answer)
>
> [15%] a) it is harder to argue that damages should be higher
>
> [55%] b) it is harder to argue that damages should be lower
>
> [30%] c) the positions are equally hard to argue for

The students were expressly told not to begin the second task (assessing the comparative difficulty question) until after they had completed the first (making arguments one way or the other).[61]

The results confirmed our hypothesis: A clear majority (55%) thought that arguing for a lower award would be the more difficult rhetorical position. Further, of those who showed a preference, the margin was almost four to one that arguing for a larger award is easier. Moreover, being asked to justify a higher or lower award had no effect; both groups agreed that it is harder to argue for a lower award. Note that the University of Chicago study closely followed the jury study, in that the former, like the latter, involved a corporate defendant with $200 million in annual profits. It seems likely, then, that a rhetorical asymmetry played a substantial role in producing jury verdicts consistently above the median individual judgment, and sometimes even above the highest individual judgment.

2. *Some Remaining Questions.* — This finding of rhetorical asymmetry raises many issues. The concept can be understood in many ways. Taken very broadly, rhetorical asymmetry is ubiquitous: In any social arrangement containing norms, those who argue in the direction that is normative will have the advantage. Those who argue that slavery was wonderful, or that the Holocaust never happened, or that animals should be made to suffer, in a context where these positions are normative, will be at a rhetorical advantage compared to those who claim the opposite. A similar rhetorical asymmetry might also be at work in other deliberative contexts. We might imagine, for example, settings in which those arguing for higher criminal punishments would have a rhetorical advantage; it is also possible to imagine places in which people arguing for lower taxes would have an easier time in any debate.

Narrow understandings of rhetorical asymmetry are possible and, for many purposes, more useful. Our claim here is quite narrow: that when people are asking "how much" questions in deliberating about punitive damage awards in dollars, one side has a systematic advantage, *even if the underlying moral judgments are identical* (as measured on the bounded punishment scale). Would the same effect be found in deliberations about compensatory awards for libel, sexual harassment, and pain and suffering? Would the same effect be found for punitive damage awards if the defendant were not a corporation? If the plaintiff were a corporation?

---

61. For a full reproduction of the questions used in this study, see infra Appendix B.

We turn to some normative issues below;[62] for the moment we note simply that there is much room for further study here.

3. *An Alternative Possibility: The Mean Juror Hypothesis.* — Another explanation for the severity shift would suggest that groups move toward the mean of individual dollar judgments, rather than the median. Because individual dollar judgments are skewed to the right, and include many extreme judgments, the mean will be above the median (this is true for 91% of juries), and could account, in theory, for the higher level of verdicts. From an analysis of our data, our basic conclusion is that while it is possible that movement toward the mean may have played some role in producing severity shifts, such movement cannot fundamentally account for them. The simplest demonstration of this comes from the fact that 27% of non-zero jury dollar verdicts were as high as or higher than that of the highest predeliberation dollar judgment of individuals. A fuller explanation requires a more detailed analysis.

To examine the hypothesis that the mean of individual awards would predict jury dollar verdicts and hence the severity shift, we recomputed the statistical jury results using the mean individual judgment, rather than the median.[63] As expected, the mean juror produces higher statistical jury dollar verdicts than the median juror, although these are still lower than 64% of non-zero deliberative jury verdicts (albeit an improvement over the 83% figure for the median juror). This partial success, however, comes at a high price. Even though the mean juror's award is consistently higher than the median juror's award (and seemingly closer to jury verdicts), the mean is less reliable, and is actually a worse predictor of jury verdicts on the conventional measures of predictive success, than is the median: It explains less of the variance in jury verdicts (4% vs. 26%), and has larger prediction errors on average (compared to the median juror predictions, the root-mean-square error[64] for the mean juror is 2.21 times larger, and the mean absolute error is 1.53 times larger). The choice between the median and the mean is mainly a matter of choosing between types of errors—with the median juror, the statistical jury's verdict is almost always too low, but almost never disastrously wrong. With the mean juror the sigus of the errors are more balanced (2/3 too low and 1/3 too high), but there can occasionally be huge positive errors

62. See infra Part IV.

63. For punishment verdicts (for clarity, not depicted here), switching to the mean as the basis for statistical juries has little effect because of the low level of skewness in the distribution of punishment judgments.

64. The root-mean-square error (RMSE) is calculated by first computing the differences between the predicted verdict (i.e., the median or average of predeliberation jurors) and the actual verdict (the jury verdict), taking the square of each difference, and then taking the square root of the average squared difference. In regression, the predicted value of the regression equation plays the same role as the mean or median do here, and the RMSE is thus analogous to the standard error of a regression (the estimate of σ). The mean absolute error (MAE) is computed by taking the average of the absolute values (i.e., the magnitude, regardless of its sign) of the differences.

(i.e., mean juror far above the jury verdict).[65] Thus, even if the mean juror with its higher overall levels did fit jury verdicts better (and it does not), we would still need to account for a consistent upward movement in jury verdicts, relative to the mean juror, and for those verdicts that are at or above the maximum juror judgment, predeliberation.

B. *Choice Shifts and Group Polarization*

1. *The Data and Some Central Ideas.* — Our study shows what is conventionally called a "choice shift" with respect to punishment ratings, pivoting around the rating of "3."[66] Choice shifts of this general sort are common consequences of deliberation, and they have been found in many diverse tasks. The result is that groups often go in more extreme directions—both higher or lower on the relevant scale—than would the typical or average individual in the group. As noted, the related phenomenon of group polarization—for which we did not test here—occurs when individuals move to a more extreme position in the direction indicated by the mean of predeliberation judgments. We offer a brief summary of relevant literature.

With respect to group polarization, consider some examples from relevant experiments. (a) A group with moderately profeminist attitudes becomes more strongly profeminist after discussion.[67] (b) Citizens of France become more critical of the United States and its intentions with respect to economic aid after discussion.[68] (c) After discussion, whites predisposed to show racial prejudice offer more negative responses to the question whether white racism is responsible for conditions faced by African Americans in American cities.[69] (d) After discussion, whites predisposed not to show racial prejudice offer more positive responses to the same question.[70] Choice shifts and group polarization stem from similar mechanisms, and for punishment ratings, the pattern described above is exactly what would be predicted from the literature on choice sbifts.[71]

---

65. This is usually due to the presence of one or two extremely high individual judgments in a jury.

66. That is, when the median individual judgement is above 3, the jury verdict tends to move up; when the median individual judgement is 3 or less, the jury verdict tends to move down.

67. See David G. Myers, Discussion-Induced Attitude Polarization, 28 Hum. Rel. 699, 707–11 (1975).

68. See Brown, supra note 60, at 223–24 (describing experiment published in Serge Moscovici & Marisa Zavalloni, The Group as a Polarizer of Attitudes, 12 J. Personality & Soc. Psychol. 125 (1969)).

69. See David G. Myers & George D. Bishop, Discussion Effects on Racial Attitudes, 169 Science 778, 778–79 (1970).

70. See id.

71. There is one difference: In the usual choice shift and group polarization studies, the phenomenon is defined by reference to scales having two sides, with a "neutral" midpoint, usually defined as zero (signaling neutrality on a question or no opinion). This is the arrangement by which it makes sense to speak of initial dispositions and their

2. *Risky Shifts.* — Before 1961, conventional wisdom had been that as compared with the individuals who compose it, a group of decisionmakers—for example, a committee or board—would be likely to favor a compromise and thus to avoid risks.[72] But the relevant experiments, originally conducted by James Stoner, found otherwise; they identified what has become known as the "risky shift."[73] Deliberation tended to shift individual members in the direction of greater risk-taking (group polarization); and deliberating groups, asked to reach a unanimous decision, were generally more risk-inclined—sometimes far more risk-inclined—than the mean individual member, predeliberation (choice shift).

In Stoner's original data, subsequent researchers noticed, the largest risky shifts could be found when group members "had a quite extreme risky initial position," in the sense that the predeliberation votes were weighted toward the risky end, "whereas the item[s] that shifted a little or not at all started out near the middle of the scale."[74] Discussion among very cautious individuals would produce a significant shift toward greater caution; discussion among individuals inclined toward risk-taking would produce a significant shift toward greater risk-taking; and discussion among individuals in the middle would produce smaller shifts in the direction indicated by their original disposition. In short, "group discussion moves decisions to more extreme points *in the direction of the original inclination* . . . . which means shift to either risk or caution in the direction of the original disposition, and the size of shift increases with the degree of initial polarization."[75] Similar results have been found in many con-

---

aggravation. Our punishment ratings, by contrast, lacked an obvious neutral midpoint, and of course dollar awards have no such midpoint.

72. We draw in this and the following paragraph on Brown, supra note 60, at 200–06 (discussing the phenomenon called the "risky shift" in the context of group polarization).

73. See James A.F. Stoner, Risky and Cautious Shifts in Group Decisions: The Influence of Widely Held Values, 4 J. Experimental Soc. Psychol. 442 (1968); James A.F. Stoner, A Comparison of Individual and Group Decisions Including Risk (1961) (unpublished M.A. thesis, School of Management, Massachusetts Institute of Technology).

74. Brown, supra note 60, at 211.

75. Id. In our study, it is plausible that the requirement of unanimity pushed people further in the direction of the dominant view, an idea that might be fortified with the thought that those with outlier positions (in favor of extreme awards) would be especially likely to hold out against a compromise view, thus producing pressure toward the extremes. We are unaware, however, of any studies of choice shifts that show a difference between the outcomes produced by a unanimity rule and the outcomes produced by majority rule. Note also that numerous studies show that group polarization occurs regardless of the decision rule and hence it is extremely unlikely that the unanimity rule accounted for our results here: "The shift effect is about equally robust regardless of whether a group decision is required." Myers & Lamm, supra note 3, at 611. Of course we cannot exclude the possibility that the results would be somewhat different without a unanimity rule; this is in fact a good area for subsequent empirical study, especially in light of continuing questions about the consequences of requirements of jury unanimity.

texts, involving, for example, questions about economic aid, political leaders, race,[76] feminism,[77] and judgments of guilt or innocence.[78]

We do not know if our jurors were susceptible to a group polarization, because members were not polled individually afterwards. But group polarization usually occurs where choice shifts are found, and hence there is reason to suspect that this happened here as well.[79] What is most noteworthy is the finding of a choice shift for punishment ratings, and the related finding of much larger severity shifts for high dollar awards than for low dollar awards.

3. *Two Mechanisms, and Severity Shifts Again.* — There have been two main explanations for group polarization and choice shifts, both of which have been extensively investigated.[80] Massive support has been found on behalf of both explanations.[81]

The first explanation involves *social comparison.*[82] On this view, people want to be perceived favorably by other group members (and also to perceive themselves favorably), and once they hear what others believe, they adjust their positions in the direction of the dominant position. They may want to sigual, for example, that they are not cowardly or cautious, and hence they will frame their position so that they do not appear such in comparison to other group members.[83] The dynamic behind the social comparison explanation is that most people may want to take a position of a certain socially preferred sort, and no one can know what such a position would be until the positions of others are revealed.[84] Thus individuals move their judgments in order to preserve their image to others and their image to themselves. This dynamic helps explain a shift toward caution (the "cautious shift") as well as toward risk-taking (the "risky shift").[85]

---

76. See Brown, supra note 60, at 224.

77. See Myers, supra note 67, at 707–12.

78. See David G. Myers & Martin F. Kaplan, Group-Induced Polarization in Simulated Juries, 2 Personality & Soc. Psychol. Bull. 63 (1976).

79. See Brown, supra note 60, at 226.

80. Brown, supra note 60, at 212–17, and Isenberg, supra note 3, review this literature.

81. Note that conformity does not explain group polarization and choice shifts. See Brown, supra note 60, at 207–08 (rejecting conformity explanation on the basis that "the risky shift is not convergence to the mean of initial positions, but, rather, to points on one side of the mean, the riskier side").

82. There is an obvious connection between the social comparison explanation and recent work in economics on reputational influences on behavior. See, e.g., Timur Kuran, Private Truths, Public Lies (1996).

83. On signaling generally, see Eric Posner, Law and Social Norms (forthcoming June 2000) (manuscript at 18–27, on file with the *Columbia Law Review*).

84. "Once the real location of the mean was known, should it not be the case, granting that everyone wanted to see himself as reasonably audacious, that those who were really below the mean would be motivated to adopt riskier positions and so change the mean and produce the risky shift?" Brown, supra note 60, at 214.

85. Investigations of social influence have emphasized both one-upmanship and the removal of pluralistic ignorance, that is, ignorance of what other people think (or are

The second explanation emphasizes the role of *persuasive arguments*.[86] The key point here is that an individual's choice or position on an issue is a function of the number and persuasiveness of arguments presented. Because a group that is inclined in a certain direction will have a disproportionate number of arguments supporting that direction, the result of discussion will be to move individuals further in the direction of their initial inclinations. Thus it is suggested that "[t]he important thing that happens in discussion is that individual arguments are expressed and become fully shared."[87] Once the set of individual arguments is exposed to all individual members, there will be a movement toward a more extreme point in the direction of initial inclinations, simply because arguments in that direction have been pressed and repeated more frequently than opposing arguments.[88]

These two mechanisms help account for severity shifts as well; they help explain rhetorical asymmetry. As our University of Chicago study suggests, arguments for higher awards are more persuasive, other things being equal, than arguments for lower awards. This is exactly what is meant by rhetorical asymmetry, in the sense that certain arguments (for "sending a stronger signal") are, we hypothesize, more convincing than others (for "ensuring against overdeterrence"). In addition, social influences, given existing norms, are likely to push people toward higher awards, simply because a concern for reputation, and for self-conception, generally argues in favor of supporting higher awards in the face of conflict. People who argue for higher awards seem to want to give appropriate punishment to wrongdoing (a good thing to seem to want), whereas those who argue for lower awards seem, other things being equal, to be lenient toward wrongdoing by corporations, or solicitous to them (not a good thing to seem to be). We emphasize that these are simply descriptive points, and that social norms could be otherwise, as they apparently are in the context of criminal conviction, where leniency shifts have been observed.[89]

IV. Implications and Reforms

To know whether the dollar awards of deliberating juries are better than the dollar awards of statistical juries, it seems necessary to have a theory of appropriate punitive damage awards, and it is not our purpose

willing to say they think). Note that it is implicit in these findings that people seem to want not to conform, but to be different from others in a desirable way.

86. There is an obvious connection between the persuasive arguments explanation and recent work in economics on informational influences on behavior, and in particular, recent work on informational cascades. See, e.g., David Hirchleifer, The Blind Leading the Blind, *in* The New Economics of Human Behavior 188, 193–207 (Mariano Tommasi & Kathryn Ierulli eds., 1995).

87. Brown, supra note 60, at 219. See Sunstein, supra note 60, for a general discussion.

88. See Brown, supra note 60, at 219.

89. See supra note 2.

here to set out such a theory. In its absence, the simplest conclusion from our study is that to the extent that there is a concern about unpredictable damage awards, deliberation is not likely to alleviate that concern, and indeed is likely to aggravate it, as demonstrated by our discussion in Part II.[90]

This unpredictability would probably be lower with larger juries (say of size twelve). But while this increase would almost certainly reduce the problem somewhat—because predictability generally increases with group size[91]—there is little reason to hope that it would make a qualitative difference. In our current study, deliberating juries of six performed no more predictably than statistical juries of the same size when using the punishment scale, and were less predictable when using the dollar scale. The statistical juries in our previous study (which were the primary basis for our conclusions about unpredictable awards) contained twelve jurors. Thus, if the relationship between deliberating and statistical juries we found here holds, it is likely that deliberating juries of twelve would show less predictability for dollar awards than our statistical juries of twelve; and those juries sbowed an extremely bigb degree of unpredictability.

Unpredictability is a serious problem for jury verdicts, partly because it ensures that the similarly situated will often not be treated similarly (and thus produces unfairness for plaintiffs and defendants alike), partly because it may produce overdeterrence in risk-averse defendants (if some of the awards are sufficiently high), and partly because of the sheer cost involved in litigation-related expenses. Of course, predictable awards might be nothing to celebrate if they are too bigh or too low. But unpredictability is in itself a cause for serious concern. How do our findings here bear on possible reforms?

With respect to damage awards, compensatory as well as punitive, many proposals have been motivated by a desire to decrease unpredictability. This goal has, for example, played a role in proposals for damage caps, for simple multipliers (relating punitive awards to compensatory awards), and for informing the jury of average awards or of intervals.[92] It has also played a role in constitutional limitations.[93] But many of these proposals would do nothing about the problem of scaling without a modulus; damage caps, for example, would reduce unduly high awards, but would not inform the jury of the meaning of various possible awards.

---

90. Note also that because our study stipulated compensatory damages, and held them fixed across cases, we may well have understated true variance in punitive awards, because according to previous research, real juries anchor on their own compensatory award, rather than on some constant value. See Eisenberg et al., supra note 10, at 637–39, 647; Karpoff & Lott, supra note 28, at 543.

91. See Saks, Tort Litigation System, supra note 43, at 1269, 1271–74.

92. See Saks et al., supra note 40, at 246 (discussing alternative approaches and their effects).

93. See BMW of North America, Inc. v. Gore, 517 U.S. 559, 585–86 (1996) (Breyer, J., concurring) (suggesting that punitive damages award is "grossly excessive" and beyond the "constitutional limit").

In fact there is evidence that caps can act as "anchors," drawing jurors to them, and hence that caps can actually increase unpredictability.[94] For these reasons damage caps are unlikely to resolve the fundamental problem.

An understanding of that problem motivated our earlier discussion of the possibility of eliciting from the jury not dollar awards, but normative judgments on a rating scale.[95] These judgments might be converted into a dollar award through some kind of calibration function, based on experts ("technocratic populism") or on population-wide data relating normative judgments to dollar awards ("predictable populism").[96] Before our study here, it would be possible to question whether it is practical to ask a deliberating jury to make a moral judgment on a rating scale, hardly an ordinary practice in daily life, and indeed a task that might seem even odder than the somewhat more familiar one of punishing wrongdoers through dollar awards.

The findings here do not lead directly to any particular reform proposal, but they add several points to the existing literature. First, they demonstrate that juries can use a punishment rating scale quite reliably. Juries are able to answer the normative question directly, and they are also able to use a rating scale far more reliably than the familiar dollar scale. And if deliberating juries are thought to have advantages over other, less populist institutions—as many people clearly believe[97]—then there is reason to consider a reform proposal that would involve directly eliciting the jury's moral judgment. As noted, this judgment might be converted into a dollar award by some kind of calibration formula, defined by expert judgments about what different dollar awards would mean or do to particular defendants, or instead by population-wide data relating normative judgments to dollar awards. Either route could greatly diminish unpredictability.[98]

The data here, along with previous data, show that a calibration formula is also feasible to develop and use.[99] In such a reform, juries might be asked to perform two simple tasks: decide whether punitive damages should be awarded at all, and produce a "punishment rating" on a scale that has verbal descriptions to accompany the numbers. It is easy

---

94. See Jennifer K. Robbenolt & Christina A. Studebaker, Anchoring in the Courtroom: The Effects of Caps on Punitive Damages, 23 Law & Hum. Behav. 353, 367 (1999) ("[W]hen the cap was high, the size and variability of the punitive damage awards were higher than awards in a control condition in which the cap was absent."). But see Linda Babcock & Greg Pogarsky, Damage Caps and Settlement: A Behavioral Approach, 28 J. Legal Stud. 341, 368 (1999) (finding "strong evidence that a cap reduces uncertainty about the trial outcome").

95. See Sunstein, Kahneman, & Schkade, supra note 4, at 2112–21.

96. See id.

97. See, e.g., Galanter & Luban, supra note 19, at 1439 (arguing that "a jury is especially suited to send the community's 'message' through the medium of damages").

98. See Sunstein, Kahneman, & Schkade, supra note 4, at 2112–20.

99. See id. at 2112–18.

to imagine a possible jury instruction that would set forth these tasks. We have shown that this approach is entirely feasible, and also that it would increase predictability.[100]

An additional point stems from the finding that deliberating juries do not reduce but actually increase erratic awards. We have seen as well that deliberation can produce juries in which group discussion yields awards much higher than those of even the highest of predeliberation judgments. Without saying whether the resulting judgments are good or bad, our findings fortify the suggestion that difficulties with the dollar scale make it hazardous to continue to rely on the current system, in which juries must map their moral judgments onto that scale without being given any guidance about the meaning of the various "points" on the scale.[101]

The most radical reform would be to dispense with the jury entirely and to move toward judicial judgments or even to develop a kind of penalty schedule, based on the judgments of some combination of representative and expert institutions.[102] We cannot evaluate these alternatives here. Of course the radical reform might be rejected if the relevant institutions would be unreliable, perhaps because bureaucracies might be vulnerable to the influence of politically powerful private groups. The question is one of comparative institutional competence. What we have added here is that the process of deliberation will increase awards generally and high awards dramatically, a result that cannot be comforting in light of the resulting unpredictability.

---

100. To be sure, eliciting moral judgments rather than dollar awards would not answer all of the relevant questions, because choice shifts produce not only higher and lower dollar awards, but also higher and lower moral judgments.

101. Cf. Lane v. Hughes Aircraft Co., 93 Cal. Rptr. 2d 60, 72 (2000) (Brown, J., concurring in the result) (citation omitted):

> [J]ury determinations of punitive damages are likely to contain an element of arbitrariness as long as the awards remain uncalibrated. To assure basic fairness, courts must consider ways of calibrating punitive damage awards. We must ask what anchoring variable will make an award of punitive damages an appropriate measure of punishment rather than a test of a jury's ability to imagine big numbers.

102. The idea has received considerable attention in the analogous area of contingent valuation. See Murray B. Rutherford et al., Assessing Environmental Losses: Judgments of Importance and Damage Schedules, 22 Harv. Envtl. L. Rev. 51, 51–56 (1998); Richard B. Stewart, Liability for Natural Resource Injury: Beyond Tort, *in* Analyzing Superfund 219, 241–44 (Richard L. Revesz & Richard B. Stewart eds., 1995). In the area of compensatory damages, see the plea for damages schedules in Bovbjerg et al., supra note 25, at 922, 937. In the punitive damage context, see Sunstein, Kahneman, & Schkade, supra note 4, at 2121–25; Viscusi, supra note 18, at 589–90. For problems with the current damages regime generally, see Atiyah, supra note 27, at 66–71.

V. Brief General Notes on Deliberation

The topic of deliberation has attracted a great deal of recent interest in both political and legal theory.[103] Much of the relevant work depends on claims about the consequences of deliberation. But for the most part, the discussion in law and political theory has not been empirically informed.[104] Our study here provides a remarkable set of data about the effects of deliberation with respect to both "pure" moral judgments (as measured by punishment ratings) and dollar awards.[105] An obvious question is whether our analysis of the data suggests that deliberation has moved people in better or worse directions. We offer some brief notations.

It might seem tempting to say that with respect to both punitive intent and damage awards, there is no basis for choosing between the results of jury deliberation and the results that would be produced by taking the median of nondeliberating six-person groups. Consider just one case, in which the median individual predeliberation punishment rating was 5.5 and the median dollar award $250,000; after deliberation, the jury rating was 7.0 and the dollar award $3 million. Which is better? Without a substantive theory about appropriate punishment ratings or dollar awards, it might seem impossible to say.[106]

If social influence and persuasive arguments are at work, there is a tendency to move to a more extreme point in the direction of the group's initial inclination; but do these mechanisms produce improvements? It seems hard to know without evaluating the initial positions that produce social influence, and without knowing whether the arguments found to be persuasive are actually right. The finding of severity shifts, apparently rooted in rhetorical asymmetry, also raises troubling questions. The argument that "a stronger signal needs to be given to other wrongdoers," or that "we need to get the attention of this defendant," seems to be far more powerful than the argument, "wait, there is a threat of overdeter-

---

103. See, e.g., Deliberative Democracy (Jon Elster ed., 1998) (collecting theories of democracy revolving around "the transformation rather than simply the aggregation of preferences"); Gutmann & Thompson, supra note 17 (developing "a conception of democracy that secures a central place for moral discussion in political life").

104. An exception is James Fishkin's set of studies of the "deliberative opinion poll." See James S. Fishkin, The Voice of the People 161–76 (1998).

105. Note, however, that our study did not and could not guarantee what many theorists of deliberative democracy take to be the preconditions of well-functioning deliberation: an absence of strategic behavior, a willingness to listen, a norm of reciprocity, and equality among members. Undoubtedly some of those in the groups we studied behaved strategically and some were not willing to listen; some were undoubtedly more equal than others. Ours is a test of real-world deliberation, not ideal deliberation. An examination of the tapes gives an overall impression, however, that the participants generally listened well and obeyed a principle of equality.

106. A more extended treatment of the normative issues raised by choice shifts and group polarization can be found in Sunstein, Kahneman, & Schkade, supra note 4, at 2075–82.

rence," or, "a punitive damage award will give the plaintiff a windfall." Here too it is not clear whether the severity shifts lead to better judgments.

It is always possible that the rhetorical asymmetry is counteracting some other kind of asymmetry, or a bias in the system. If, for example, jurors have a systematic bias against personal injury actions, a pro-plaintiff rhetorical asymmetry, with respect to dollar awards, might supply a corrective. What is disturbing about the rhetorical asymmetry that we have described is not necessarily that it produces worse awards, but its mechanical, case-independent quality. If the result of the rhetorical asymmetry is to produce better awards, it would not be a shock—stranger things have happened—but it would be a lucky coincidence.

To be sure, there may be procedural reasons to have some confidence in the outcomes of a deliberative process. By hypothesis, more time is spent on the prohlem, and more time might well help, at least in general—judgments reached after deliberation will be more informed, simply because more arguments will be introduced; deliberation tends to increase consensus; and deliberation will, other things being equal, produce movement in the direction favored by more confident group members, and more confident people are likely (though hardly certain) to have some reason for their confidence. On procedural grounds, these points give some reason to think that with respect to punishment ratings, the outcomes of deliberation are likely to be better than the outcomes that would be produced by identifying the median or mean judgment of individuals. But because of the arbitrariness introduced by the selection of a modulus, we have no such confidence for dollar verdicts. There is little reason to believe that the dollar awards of actual juries are better than the dollar awards of statistical juries. If all existing punitive damage awards were douhled, or subject to a sliding scale of increase, so that small awards would go up a little, and large awards would go up a lot, would the system of civil penalties be better? We cannot insist on a negative answer, but it is far from obvious how one would defend an affirmative answer.

CONCLUSION

We have found that as compared with the median of individual judgments, deliberation makes low punishment judgments decrease and high punishment judgments increase. It also makes—and this is our most important finding—dollar awards generally increase, while making high dollar awards substantially increase, in a general severity shift. We have also found, somewhat to our surprise, that deliberating juries produce more unpredictability than would be found by taking the median of jurors' predeliberation judgments.

These findings have implications for damage awards in general and also for understanding social deliberation. From the normative point of view, it is hard to know whether deliberative verdicts are better than the

median of predeliberative individual judgments. But five points seem clear. First, moral judgments about personal injury cases are very widely shared over diverse communities and demographic categories. Second, those shared moral judgments do not produce predictable dollar awards. Third, choice shifts occur in the context of punishment ratings; hence group judgments go to more extreme points in the direction of the inclination originally indicated by the median of predeliberation judgments. Fourth, dollar awards reflect a systematic severity shift, apparently a result of a rhetorical asymmetry in which arguments for higher awards have a general advantage over arguments for lower awards. Fifth, the problem of unpredictable and erratic dollar verdicts is increased, not alleviated, by the fact that juries are deliberative bodies.

APPENDIX A: *GLOVER V. GENERAL ASSISTANCE*

Joan Glover, a five-year-old child, ingested a large amount of a non-prescription allergy medicine called Allerfree, and required a three-week hospital stay. The Allerfree bottle used a faulty childproof safety cap. The Glovers sued the manufacturer of Allerfree, the General Assistance company. The trial jury ordered General Assistance to pay the Glovers $200,000 in compensatory damages.

Facts of the Case Established at Trial. — Joan's parents testified that after her birth they had "childproofed" their house and ensured that all of their medications had childproof safety caps. The Allerfree bottle carries a label reading "Childproof Cap." Joan found the pills in a kitchen drawer and ingested most of the bottle. The overdose permanently weakened her respiratory system, which will make her more susceptible to breathing-related diseases such as asthma and emphysema for the rest of her life.

General Assistance is a large company (with profits of $100–200 million per year) that manufactures a variety of non-prescription medicines. The company has sold tens of thousands of bottles of medicines with childproof safety caps that were generally effective, but had a failure rate much higher than any others in the industry. Internal company documents showed that General Assistance had chosen to ignore federal regulations requiring more effective safety caps. An internal memo presented at trial says that "this stupid, unnecessary federal regulation is a waste of our money"; it acknowledges the risk that Allerfree might be punished for violating the regulation but says "the punishments are extremely mild; basically we'd be asked to improve the safety caps in the future." An official at the Food and Drug Administration had previously warned a General Assistance executive that the company was "on shaky ground on this one."

Closing Argument by Glovers' Attorney. — The attorney for the Glovers argued that General Assistance's disregard for children's safety and for the law was abhorrent and represented exactly the kind of reckless corporate greed deserving of a high award of punitive damages. He concluded that General Assistance's shocking profit-mongering should be punished so that the company would not feel itself at liberty to put children at risk in the future.

Closing Argument by General Assistance's Attorney. — The attorney for General Assistance emphasized that while the cap had a high failure rate relative to others on the market, it had nonetheless been conceded at trial that the cap was effective in most cases. She argued that, given that the FDA official had only communicated with General Assistance verbally, and had not required the company to take any action, it was not at all clear that the cap was actually in violation of the regnlation at all.

APPENDIX B: RHETORICAL ASYMMETRY QUESTIONNAIRE

A Jury's Decision

Imagine that a jury in a civil trial is deliberating about a personal injury case in which the defendant is a large corporation (with annual profits of approximately $200 million). The jury has already (a) unanimously ordered the defendant to pay an amount of compensatory damages that fully compensates the plaintiff, and (b) unanimously concluded that while the underlying conduct was not truly horrendous, it was sufficiently reckless to justify an award of punitive damages as well (in addition to compensatory damages).

The jurors then decided to think individually about the proper amount of punitive damages, prior to deliberating as a group.

They have now made their individual determinations, and group deliberation has begun.

Please do not turn to the next page until requested to do so.

Deliberating over an Amount of Punitive Damages

Juror A states an amount of punitive damages. Juror B states an amount that is much higher, three times A's amount. (You are not expected to know the exact amounts stated by A and B.)

What arguments would you expect to hear from Juror B, to support an award that is higher than Juror A's proposal?

On the lines below, please write down a list of the arguments that you might expect to hear from Juror B in support of a higher award. Write as many distinct arguments as you can, briefly summarizing each idea in a short phrase or sentence. Please write only one distinct idea per line.

Which is Harder to Defend?

In general, which position would you expect to be harder for a juror to argue for in a deliberation? (please circle the letter of your answer)

a) it is harder to argue that damages should be *higher*

b) it is harder to argue that damages should be *lower*

c) the positions are equally hard to argue for

Very briefly, please write down the reason for your answer in the space below.