Journal of Risk and Uncertainty, 16:49–86 (1998)

Shared Outrage and Erratic Awards: The Psychology of Punitive Damages

DANIEL KAHNEMAN
Princeton University, Woodrow Wilson School of Public Policy, Princeton, NJ 08544

DAVID SCHKADE
University of Texas, Austin

CASS R. SUNSTEIN
University of Chicago

Abstract

An experimental study of punitive damage awards in personal injury cases was conducted, using jury-eligible respondents. There was substantial consensus on judgments of the outrageousness of a defendant's actions and of the appropriate severity of punishment. Judgments of dollar awards made by individuals and synthetic juries were much more erratic. These results are familiar characteristics of judgments made on unbounded magnitude scales. The degree of harm suffered by the plaintiff and the size of the firm had a pronounced effect on awards. Some judgmental tasks are far easier than others for juries to perform, and reform possibilities should exploit this fact.

Key words: Punitive damages, law and psychology, jury decision making

JEL Classification: K00—Law and Economics: General

A system of justice must make distinctions: it should treat alike cases that are relevantly alike, and it should treat differently cases that are relevantly different. A system that metes out erratic and unpredictable punishments for similar transgressions fails to satisfy the first of these criteria. A system that provides crude punishments poorly linked to the distinctive features of particular cases fails to satisfy the second. Many critics have charged that the system of punitive damages is prone to such failures (e.g., Jeffries, 1986; Huber, 1989).

We use the terms 'erratic' and 'unpredictable' interchangeably in this article, both in a rather unusual sense: we imagine a legal expert trying to anticipate the award that will be made in a particular case, either by considering similar cases in the past, or by examining the responses of a mock jury. The confidence that the expert may reasonably attach to her opinion about the outcome of the case depends on how much the decisions of different juries considering the same case are likely to differ. If they do not treat the similarly situated similarly, awards that are in this sense erratic or unpredictable are unjust; they are also inefficient from the economic point of view, because they impose an unnecessary

burden of uncertainty both on plaintiffs (and their lawyers) and on potential defendants. If the social function of punitive damages is to provide a deterrent signal, as many theorists argue (see Polinsky and Shavell, 1997), unpredictability reduces the quality of that signal. Erratic awards are an ill wind that blows no one any good.

The awarding of punitive damages is a decision, and like other decisions it can be examined usefully from three distinctive points of view: normative, descriptive, and prescriptive (Bell, Raiffa and Tversky, 1988). Normative analysis is concerned with what ought to be. A normative analysis of punitive damages will seek to anchor this particular aspect of legal practice in a broader view of justice and of the purpose of the legal system. Existing normative treatments of the issue come in several flavors, which variously emphasize ethical concerns, most prominently retribution, and economic efficiency (Galanter and Luban, 1993; Polinsky and Shavell, 1997; Landes and Posner, 1993). A descriptive analysis seeks an understanding of the factors that govern actual decisions. A descriptive treatment of punitive damages will be concerned with an analysis of the social and psychological factors that affect the decisions of juries. Finally, a prescriptive analysis considers ways to overcome the limitations of decision makers, and thereby to improve decisions. The goal of a prescriptive approach to punitive damages will be to change the jurors' task so as to make it easier for them to reach normatively appropriate decisions.

The main focus of the present article is descriptive. We offer a psychological theory—called the **outrage model**—which describes the process by which individual jurors set punitive awards. We also identify a stage of this process that could be responsible for much of the random variability in these awards. We conclude that the critical stage is the mapping of a judgment about the appropriate severity of punishment (which we will call **punitive intent**) onto the dollar scale. By itself, punitive intent is quite predictable in the sense that we defined earlier: for the personal injury cases explored in our study, there is substantial social consensus about the severity of punishment that is appropriate for different cases, and different juries viewing the same cases will generally form similar punitive intentions. The consensus breaks down, however, when jurors are asked to express punitive intent in dollars, the response mode required by the legal system. Our evidence suggests that different juries considering the same case would often assign very different damages, even if they agree in their punitive intent.

The prescriptive task of legal reform should be informed jointly by a psychological analysis of judgments of punitive damages and by a normative assessment of the appropriate role of community sentiment in the award of such damages. Our analysis identifies three positions on this issue, ordered below from least to most radical: (i) community sentiment about awards is normatively acceptable, but juries provide a poor estimate of this sentiment, because of sampling error; (ii) the community's punitive intent is normatively acceptable, but community sentiment about dollar awards is poorly informed and lacks normative force; (iii) the community's punitive intent must be rejected as a normative standard, because it is an illegitimate basis for punitive damage awards. Naturally, these three assessments are associated with quite different suggestions for legal reform. The normative and prescriptive analyses are developed more fully in a companion paper (Sunstein, Kahneman and Schkade, 1998).

We proceed as follows. Section 1 sets out the outrage model and outlines the goals of the study. Section 2 describes the procedures and instruments used in the research. Section 3 presents the results. Section 4 reviews the argument and sketches the implications of the findings for legal reform. Section 5 concludes.

1. The outrage model

The vast literature on punitive damages has been mainly normative; it is concerned with the social function of punitive awards and with the legal criteria for granting them. Standard accounts suggest that punitive damages are justified as a means of overcoming the problem of underdeterrence created by the tort system's failure to detect and compensate all injured plaintiffs (Landes and Posner, 1993), and also as a way of reflecting the "sense of the community" about the appropriate retribution that should be imposed on especially serious wrongdoers (Galanter and Luban, 1993). There is little room in most of these treatments for the emotions of outrage and indignation (for an exception, see Hampton, 1993). These emotions are of course central to a psychological analysis. Although there are some exceptions, punishment is generally best understood as an act which expresses emotions of anger and indignation and is intended to cause suffering (Hampton, 1993; Kahan and Nussbaum, 1996).

We will propose a descriptive model of the psychology of punitive awards, building on a more general theoretical analysis (Kahneman and Ritov, 1998), which extends the concept of attitude beyond the domain of public issues in which it is most often applied. There is much evidence that a basic process of evaluation goes on continuously as people respond to the objects and events of their lives (Bargh et al., 1996). As we use the term, each of these objects and events evokes an attitude, which combines an emotional evaluation and a response tendency. The evaluations range from intense liking to intense dislike, and the response tendencies vary from approach to avoidance or from support to aggression. The evaluations and intentions that define attitudes can be expressed overtly in multiple ways, ranging from physiological indications of emotion to responses in opinion surveys.

The main implication of the present analysis is that highly diverse responses will correlate closely if they express the same attitudes. Indeed, high correlations (often in excess of .80) have been observed between **group averages** of different attitude measures, computed over a set of public issues (Kahneman and Knetsch, 1992; Kahneman and Ritov, 1994, 1998; Kahneman et al, 1993). The results of the present study will confirm the generality of this observation.

Figure 1 summarizes the outrage model of punitive damages, in which these damages are considered an expression of an angry or indignant attitude toward a transgressor. The evaluative aspect of the attitude is labeled **outrage**. The response tendency is labeled **punitive intent**. The goal of the model is to account for differences in the average responses of populations of respondents to various actions and various defendants. The figure is to be read from left to right.

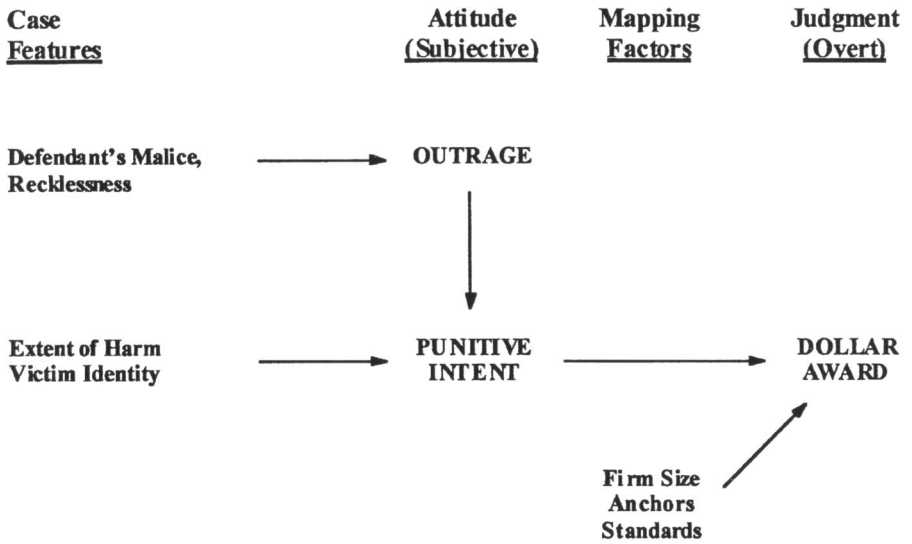

Figure 1. The Outrage Model of Punitive Damages

The variables listed under 'case features' in Figure 1 are some of the characteristics of an action that determine the intensity of outrage and the severity of punitive intent. The figure represents outrage as basic, and punitive intent as determined jointly by outrage and by additional factors. This representation is appropriate because the determinants of outrage are a proper subset of the case features that control punitive intent. Two factors that affect punitive intent independently of outrage are identified in the figure: the harm suffered by the victim and the relationship between the victim and the individual making the judgment.

The description of an action can evoke outrage without any information about the harm that it caused. However, the harm that an action caused is a factor in punitive intent. This retributive aspect of punishment is incorporated in many aspects of the law, such as the large discrepancy between the sentences for murder and for attempted murder. The relationship between victim and juror was manipulated in an experiment by Hastie, Schkade and Payne (1997); awards were larger when the plaintiff was located in the jurors' community than when the plaintiff was from a remote location. We speculate that the retributive urge is stronger when the victim belongs to one's group than when the victim is a stranger. Although punitive intent is affected by some factors that do not affect outrage, our general hypothesis is that most of the determinants of the two states are shared—and we therefore expect them to be highly correlated.

In some situations the expression of an attitude is not spontaneous, but is restricted to a particular scale of responses. Stringent constraints apply, for example, when respondents in opinion surveys are asked to indicate the extent of their agreement or disagreement with a particular statement. In the situation with which we are concerned here, the responses of juries dealing with punitive awards are restricted to a scale of dollars. The

different expressions of an attitude are affected by specific factors that may induce some (usually slight) differences among them. Thus, the column labeled 'mapping factors' in Figure 1 identifies some variables that affect the mapping of punitive intent onto the dollar scale, but not onto a rating scale of punitive intent. For example, we propose that the size of the defendant firm is a scaling factor in translating punitive intent into dollars: a judgment that appears severe when the defendant is a small firm may appear grossly inadequate when the defendant is a giant. Other factors that affect the accuracy of the mapping of punitive intent onto the dollar scale are discussed in the following section.

In summary, the outrage model assumes internal states of outrage and punitive intent, which can be mapped onto different response scales. As we see next, these scales vary both in their complexity and in the precision and reliability of the measurement that they support: some scales are 'noisier' than others. The main conclusion of this paper will be that the dollar scale is an extremely noisy expression of punitive intent.

1.1. Mapping psychological states into overt responses

For more than a century, psychologists have studied the processes by which internal states are mapped onto overt responses. An important distinction drawn in this research is between category scales and magnitude scales. Category scales consist of a bounded set of ordered responses, as in the familiar format of many opinion surveys. The categories can be represented by numbers; in such cases descriptive labels are always attached to the extremes of the scale, and sometimes to some or all intermediate values. It is generally assumed that category scales are invariant to a linear transformation: a response of 5 on a scale that ranges from 1 to 7 is interpreted as equivalent to a response of 40 on a scale that ranges from 0 to 60. As this example illustrates, the zero point on a category scale has no particular significance.

Internal states can also be mapped onto magnitude scales, which have a meaningful zero and no upper bound. The study of magnitude scaling was pioneered by S.S. Stevens (1975). The standard application of this technique is in the measurement of the psychophysical functions that relate the intensity of subjective experiences to the intensity of the physical stimulus that evokes them. In a typical experiment, the observer is exposed to a series of sounds that vary in loudness, or lights that vary in brightness, or electric shocks that vary in painful intensity. The observer is instructed to assign a positive number to each stimulus according to the relative intensity of the subjective experience that it evokes. Magnitude estimation has also been applied to social judgments, such as the heinousness of crimes and the severity of punishments (Stevens, 1975; Lodge, 1981).

The research on magnitude scaling has yielded several robust and quite general findings (Stevens, 1975). (1) There is substantial agreement among observers on the ratios of the magnitudes that they assign to the internal responses evoked by particular stimuli. For example, if one observer assigns three objects magnitudes of 1, 2 and 6, another observer who assigns a value of 5 to the first stimulus would be expected to assign values of 10 and 30 to the other two. (2) The distribution of judgments of any particular stimulus is positively skewed; skewed distributions are observed both in an individual's judgments of

the same stimulus over repeated occasions and for the judgments of a stimulus by different judges. (3) The standard deviation of the judgments of a stimulus is proportional to the mean of these judgments; this relationship also holds both for an individual's judgments of the same stimulus over repeated occasions and for the judgments of a stimulus by different judges. (4) Unless the experimenter specifies a standard stimulus and a rating that should be assigned to it (called a modulus), there are large differences in the mean values of judgments: some observers assign generally high numbers to all stimuli, others assign low numbers. (5) In general, magnitude judgments of sensory intensity are related by a power function to the relevant physical variable: for example, brightness is a power function of luminance and loudness is a power function of sound amplitude (both with an exponent of approximately 1/3). (6) When the same stimuli are independently scaled on a magnitude scale and on an interval scale, the values of stimuli on the magnitude scale are related by a power function to their value on the category scale.

Like other internal states that vary in intensity, outrage and punitive intent can be expressed either on category scales or on magnitude scales. For example, respondents could be asked to evaluate the outrageousness of different actions on a category scale ranging from "not outrageous" to "extremely outrageous". Alternatively, respondents could be asked to rate outrage on an unbounded numerical scale, where zero indicates no outrage, and a particular number (the modulus) is assigned to the outrageousness of a specified action (the standard). The modulus defines the unit of measurement. We would expect to observe characteristic differences between the judgments on the two scales. In particular, the standard deviations of the judgments of different cases should be very strongly correlated with the means of judgments expressed on the magnitude scale, but not on the category scale.[1]

The dollar scale of punitive awards satisfies the technical definition of a magnitude scale of punitive intent: the scale has a meaningful zero and no upper bound. The unit of measurement (dollars) is specified in this case. Our main hypothesis, however, is that the unit in which judgments are denominated is largely meaningless to potential jurors, who lack relevant experience in mapping their punitive intent onto the dollar scale. We therefore expect judgments of punitive awards to exhibit the high variability and positive skew that are observed in magnitude scaling without a modulus: individuals who agree in their underlying punitive intent may differ widely in the overall level of awards they assign, even if they agree on the ratios of awards for different cases. To test this prediction we compare expressions of punitive intent in two formats: category-scale ratings of the appropriate severity of punishment and judgments of punitive awards in dollars.

We use the category scales of outrage and of punitive intent, and the dollar scale of punitive damages to examine two related hypotheses. The first is that outrage and intent to punish conform to a bedrock of socially determined norms. The mapping of attitudes onto category scales with verbally defined categories provides sufficient guidance to achieve some uniformity in the use in the scales, and therefore preserves the underlying social consensus. In contrast, there are few constraints on the mapping of punitive intent to dollars. Our second hypothesis is that large individual differences in the use of the dollar scale induce extremely large variances when jurors are asked to express their punitive intent in dollar amounts. Here is the critical point: Because of the unfortunate

features of the dollar scale, two juries that share the same punitive intentions may set very different punitive awards.

1.2. The study

We designed a study to test several implications of the outrage model. Respondents were given a set of vignettes of personal injury cases in which a plaintiff (always an individual) sued a firm for compensatory and punitive damages. The respondents were told to assume in all cases that compensatory damages had been awarded in the amount of $200,000, and that punitive damages were to be considered. Different groups provided different judgments of these scenarios. Respondents in the outrage condition were asked to rate the outrageousness of the defendant's behavior. Respondents in the punishment condition rated how severely they wished to punish the defendant. Finally, respondents in the dollar award condition were asked to assess the amount of punitive damages (if any) that should be awarded.

We examined the following hypotheses:

(1) **The outrage evoked by scenarios of tortious behavior is governed by broadly shared social norms**. The ranking of different scenarios by their outrageousness is expected to be generally similar for different demographic groups (See Section 3.1.3).
(2) **Punitive intent is determined by the outrageousness of the defendant's behavior and by other case factors, prominently including the harm suffered by the plaintiff**. The prediction that harm affects punitive intent but not outrage was tested by presenting alternative versions of some scenarios, in which the severity of the harm suffered by the plaintiff was varied (See Section 3.1.4).
(3) **Damage awards are determined by punitive intent and by specific mapping factors, prominently including the size of the defendant firm**. This prediction was tested by presenting each scenario in two versions, in which the size of the defendant was varied (See Section 3.1.5).
(4) **Large individual differences in the use of the dollar scale reduce the precision with which population norms for punitive awards can be estimated from the judgments of small samples of respondents, such as juries**. This hypothesis was tested in two ways. First, we compared the amount of measurement error in ratings of punitive intent on a category scale and in dollar awards. Second, we predicted that transformations of dollar awards that remove individual differences in scale usage (e.g, transforming each individual's responses to ranks) would result in a sharp reduction in unsystematic variability. The rationale for this prediction is that such transformations eliminate the skewness and reduce the arbitrary variability associated with the dollar scale but retain the ordinal correspondence of punishments to punitive intentions (See Sections 3.2 and 3.3.1).
(5) **Because there is less error variance in ratings of punitive intent than in dollar awards, these ratings are a more accurate predictor of community sentiment about appropriate dollar awards than dollar awards themselves**. The dollar award

that a jury makes in a particular case can be thought of as an estimate of community sentiment about the appropriate award, which we define as the median of individual judgments in the population. It is possible in principle that jury ratings of punitive intent could provide a more accurate estimate. We tested this hypothesis by comparing the precision of estimates based on judgments of dollar awards and of punishment ratings (See Section 3.3.2).

2. Method

2.1. Sample and procedure

The sample consisted of 899 jury-eligible respondents who were recruited from the Travis County, Texas voter registration list by a professional survey firm and paid $35 for their participation. The resulting sample had good representation from various demographic and socio-economic groups. For example, respondents were 44% male; 64% Caucasian, 16% African-American, 15% Hispanic; median income = $30K–$50K; median education = Some College; median age = 30–39. Thirty-two respondents were dropped from the sample because they gave incomplete responses or failed to understand the task.

The survey was conducted at a downtown hotel. Participants were run in large groups at pre-arranged times over a four day period. Most respondents completed their task in 30 to 45 minutes.

Each respondent received three pages of general instructions and four numbered envelopes. The first three envelopes contained the materials for Parts 1, 2 and 3 of the study, as described below. The fourth envelope contained demographic questions and debriefing information. The instructions (which are excerpted in Appendix I) included (1) an overview of the survey procedure, (2) an explanation of the task of jurors in civil (as opposed to criminal) trials, (3) definitions of and distinctions between compensatory and punitive damages, including the fact that compensatory damages had already been awarded in the cases they would consider, (4) a summary of standard legal conditions for punitive damages (maliciousness or reckless disregard for the welfare of others), and (5) a reminder about the standard of evidence required in this situation (preponderance of the evidence).

2.2. Design and stimuli

Ten scenarios describing personal injury cases were constructed (summarized in Table 1; full versions in Appendix II). The first six were used in Parts 1 and 2 of the procedure and the other four in Part 3. Each respondent rated some version of all ten scenarios. Envelope #1 contained material about one of the first six scenarios. Envelope #2 contained the other five of these six scenarios. Envelope #3 contained the four scenarios used in Part 3 of the experiment.

In Parts 1 and 2, between-subjects manipulations were response mode (Outrage, Punishment or $ Damages) and firm size (annual profits of $10–$20 million (Medium) or

Table 1. Summary of Personal Injury Scenarios

Case	Part	Description
Mary	1,2	Employee suffers anemia due to benzene exposure on the job
Frank	1,2	Motorcycle driver injured when brakes fail
Thomas	1,2	Circus patron shot in arm by drunk security guard
Susan	1,2	Auto airbag unexpectedly opens, injuring driver
Carl	1,2	Man suffers skin damage from using baldness cure
Sarah	1,2	Elderly woman suffers back injuries from using exercise video
Jack	3	Small child playing with matches burned when pajamas catch fire
Joan	3	Child ingests large quantity of allergy medicine, needs hospital stay
Martin	3	Disabled man injured when wheelchair lift malfunctions
Janet	3	Secretary chronically ill due to radiation from computer monitor

$100–$200 million (Large)), and scenario sequence, including which scenario was evaluated first, in isolation from the others (in Part 1 of the procedure). Scenario order was counterbalanced so that each scenario appeared in each ordinal position with equal frequency. Table 2 shows the wording of the evaluation questions in the three response modes. Instructions in all scenarios stated that compensatory damages of $200,000 had already been awarded.

Part 3 had the same structure as Parts 1 and 2, except that the isolation manipulation was replaced by a manipulation in which the degree of harm suffered by the plaintiff was varied. For each of the four scenarios used in Part 3, we formulated a high-harm and a low-harm version. For example, in the case in which a child playing with matches was burned when his pajamas caught fire, the injuries were described as "He was severely

Table 2. Response Mode Manipulation

Outrage

Which of the following best expresses your opinion of the defendant's actions? (please circle your answer)

Completely Acceptable		Objectionable		Shocking		Absolutely Outrageous
0	1	2	3	4	5	6

Punishment

In addition to paying compensatory damages, how much should the defendant be punished? Please circle the number that best expresses your opinion of the appropriate level of punishment.

No Punishment		Mild Punishment		Severe Punishment		Extremely Severe Punishment
0	1	2	3	4	5	6

$ Damages

In addition to paying compensatory 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? (please write your answer in the blank below)

$ _____

burned over a significant portion of his body and required several weeks in the hospital and months of physical therapy" (high harm) or "His hands and arms were badly burned, and required regular professional medical treatment for several weeks" (low harm).

3. Results

We discuss the results in three parts. Section 3.1 is concerned with the validity of the outrage model shown in Figure 1. Section 3.2 deals with the predictability of dollar damage awards, and with the role of individual differences. Section 3.3 considers the question of unpredictability at the level of juries. We examine the hypothesis that variability in the individuals' use of the dollar scale is sufficient to be a major cause of unpredictability in jury awards.

3.1. The outrage model

The analyses in this section examine some of the social norms that govern evaluations of the actions of firms and the intended severity of punishment for harmful negligence. They focus on the average judgments of the general public (as represented in our sample), and of sub-populations within it.

3.1.1. Preliminary analysis: The effect of context. Unlike real jurors, who are exposed to a single case for a long time, the participants in our study responded to a total of 10 personal injury cases in quick succession, and had an opportunity to compare these scenarios to each other. To examine the effect of this unusual procedure, every participant first encountered one of the first six scenarios in Table 1, which was presented in a separate envelope and was evaluated in isolation from the others (Part 1). The experimental design provides a comparison of the distribution of judgments to each scenario when it is judged in isolation or in the context of other scenarios. The context was provided in envelope 2, where subjects were instructed to read through all five scenarios therein before responding to any one. In addition, of course, each subject had already responded to the scenario they encountered in envelope 1, which also contributed to the context. The relevant data are summarized in Table 3, which shows the mean outrage and punishment judgments and median dollar awards for each scenario in the isolation and in the context conditions.

To test whether the availability of a context altered intuitions, we computed rank-correlations ($N = 12$) over the scenarios of Table 3, between the judgments in the isolation condition and in the context condition. The correlations are high: .88, and .90, respectively, for the means of outrage and punishment ratings, .89 for the medians of dollar awards.

We also examined whether the availability of a context of comparison affected the distribution of judgments. Except for a small but significant increase in the severity of punishment in the context condition ($t = 3.19$, $p < .01$), the means of judgments made in

Table 3. Comparison of Context and No-Context Versions of Same Scenario

	Mean Outrage		Mean Punishment		Median $ Awards	
	No Context	Context	No Context	Context	No Context	Context
Medium Firms						
Mary	3.76	4.43	3.64	4.36	500,000	600,000
Frank	3.76	4.73	4.35	4.58	550,000	500,000
Thomas	4.73	4.68	3.85	4.81	205,000	500,000
Susan	3.12	2.91	3.00	3.15	100,000	200,000
Carl	2.92	2.66	2.25	2.65	200,000	200,000
Sarah	1.92	1.24	0.75	0.59	0	0
Mean	3.37	3.44	2.97	3.36	259,167	333,333
Stdev (column)	0.95	1.41	1.31	1.60	220,000	234,000
Stdev Ratio		1.48		1.22		1.06
Large Firms						
Mary	3.68	4.13	3.72	4.66	500,000	1,000,000
Frank	4.23	4.34	4.15	4.90	600,000	1,000,000
Thomas	4.62	4.22	4.39	4.80	250,000	500,000
Susan	2.89	2.84	2.76	3.08	500,000	250,000
Carl	2.29	2.89	3.08	2.73	200,000	200,000
Sarah	2.25	1.41	1.17	0.88	0	0
Mean	3.33	3.31	3.21	3.51	341,667	491,667
Stdev (means)	1.00	1.15	1.18	1.59	229,000	425,000
Stdev Ratio		1.15		1.35		1.86

isolation and in context are quite similar. The variances of judgments within each scenario are also similar. The most consistent effect of a context of similar cases is to increase the range of the judgments of different scenarios. The variances of mean judgments of outrage and punishment over the 12 scenarios are larger in the context condition. In a test for a difference between correlated variances, the effect of context is significant for the punishment condition ($F(1,10) = 6.87, p < .05$), and marginally significant for outrage ($F = 4.07, p < .10$). The results for a similar comparison of the medians of dollar awards are similar ($F = 4.64, p < .10$). The availability of a context apparently makes a serious case appear more serious than it would on its own, and makes a milder case appear milder.

The small effect of a comparison context in improving discriminations among cases could be useful in the design of possible reforms of the jury's task. However, the fact that the effect of context is small allows us to generalize the conclusions of the present experiment with some confidence to situations in which no context is provided (the standard case for real juries) or to situations in which other scenarios might be used.

3.1.2. Average judgments of scenarios. Table 4 presents measures of central tendency for the three types of response (outrage, punishment and awards) for each of the scenarios presented in the experiment. Because the context effects analyzed in the preceding section are small, judgments obtained in Part 1 and in Part 2 of the questionnaire are pooled in the

Table 4. Aggregate Responses by Scenario and Condition

Firm Size	Harm Level	Scenario	Mean Outrage	Mean Punish	Mean $ Awards	Median $ Awards	% Zero $ Awards
Large	High	Joan	4.24	4.93	17,853,229	2,000,000	5.7
Large	High	Martin	3.88	4.78	17,071,115	1,900,000	0.0
Large	—	Frank	4.32	4.77	9,954,507	1,000,000	0.7
Medium	High	Martin	4.41	4.73	4,899,710	1,000,000	2.9
Large	—	Thomas	4.29	4.73	8,703,479	500,000	2.8
Large	Low	Joan	4.19	4.65	22,131,390	1,000,000	2.8
Medium	—	Thomas	4.69	4.64	2,152,765	525,000	2.9
Medium	—	Frank	4.56	4.54	3,450,993	500,000	1.5
Medium	High	Joan	4.49	4.53	4,871,791	1,000,000	6.0
Large	—	Mary	4.05	4.50	9,162,137	1,000,000	2.8
Medium	Low	Martin	4.03	4.32	2,185,522	525,000	1.5
Medium	Low	Joan	4.67	4.32	6,204,239	550,000	7.2
Large	Low	Martin	4.01	4.28	9,589,643	1,000,000	2.9
Medium	—	Mary	4.32	4.24	1,939,926	575,000	2.2
Medium	—	Susan	2.95	3.12	1,701,522	200,000	18.4
Large	—	Susan	2.85	3.03	10,254,317	300,000	12.7
Large	High	Janet	2.91	2.79	5,122,207	250,000	28.0
Large	—	Carl	2.80	2.79	1,779,264	200,000	9.9
Medium	—	Carl	2.70	2.59	752,173	175,000	16.2
Large	Low	Janet	2.84	2.56	7,229,776	200,000	23.9
Medium	High	Janet	2.65	2.55	1,263,264	200,000	12.5
Medium	Low	Janet	2.49	2.54	1,925,031	150,000	29.7
Medium	High	Jack	1.83	1.58	1,289,063	32,500	43.8
Large	Low	Jack	1.75	1.56	1,650,695	100	49.3
Large	High	Jack	1.83	1.50	917,836	0	58.2
Medium	Low	Jack	1.49	1.18	483,403	0	66.7
Large	—	Sarah	1.54	0.93	89,314	0	74.6
Medium	—	Sarah	1.35	0.61	230,184	0	75.7

data of Table 4. Different summary statistics are used for the different variables. Means are shown for ratings of outrage and of intended punishment. For dollar awards, the Table presents medians and the proportion of respondents who indicated that the award should be zero. The use of the median is justified because the distribution of dollar awards was severely skewed, as expected for judgments on a magnitude scale. The scenarios appear in descending order of the mean punishment rating. Separate values are shown for the medium-firm and large-firm versions of each scenario, and also for the low-harm and high-harm versions of the four scenarios used in Part 3 of the questionnaire.

The main observation from Table 4 is that the order of the scenarios from worst to mildest is quite similar for the three responses. The means of outrage ratings and dollar awards generally decline, concordant with the mean punishment ratings that were used to arrange the scenarios. We computed the Pearson correlation between the values shown in Table 4 ($N = 28$). The correlation between the means of outrage and punishment ratings is very high (.98). The correlations of these two variables with median dollar awards are lower (.69 and .78, respectively for outrage and punishment ratings). As we shall see later,

these correlations are attenuated by the low signal/noise ratio of the dollar response. Because the skewed distribution of dollar awards has a more severe effect on the stability of sample means than of medians, the correlations of outrage and punishment ratings with the means of dollar awards are even lower (.51 and .61, respectively).

3.1.3. Demographic analyses. Do the averages shown in Table 4 represent responses to broadly shared social norms? Are there different norms for different social groups? To answer these questions we broke down the sample according to different criteria, including ethnicity, gender, education and income. A table similar to Table 4 was computed separately for each subcategory (e.g., women and men). We then computed Pearson correlations between the mean judgments of mutually exclusive categories of respondents, over the 28 scenarios. Table 5 presents the correlations for mean punishment ratings.

The results of Table 5 indicate that the ratings of punitive intent reflect norms on which there is little or no disagreement between different social groups. Men and women, Hispanics and whites, and respondents at very different levels of income produce very similar orderings of the 28 scenarios used in the study. Judgments of intent to punish in these scenarios of personal injury cases evidently rest on a bedrock of moral intuitions that are broadly shared in society.

The results for outrage ratings are generally similar, though the correlations are slightly lower (a median correlation of .94, compared to .99 for punishment ratings). The lower correlations reflect the slightly lower statistical reliability of outrage ratings. The correlations between mean judgments of dollar awards are much lower (the median correlation was .65), another indication of the poor reliability of this measure. This unreliability, of course, could be mitigated by using larger samples.

Table 5. Correlation Between Demographic Groups on Intended Severity of Punishment[a]

		Men		
Gender	Women	.974		
		Black	White	
Ethnicity	White	.975		
	Hispanic	.963	.988	
		<30K	30–50K	
Household	30–50K	.991		
Income	>50K	.986	.986	
		<30	30–39	40–49
Age	30–39	.994		
	40–49	.992	.994	
	>50	.991	.993	.987

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

The high correlations shown in Table 5 imply only that the relationships between the mean judgments of disjoint demographic groups are well described by linear functions. They do not preclude systematic differences between groups in their use of the response scales. Thus, one group could be generally more punitive than another even if the two make the same discriminations among scenarios. Groups may also differ in the extent to which they discriminate between cases, yielding different variability over cases.

To test for differences between the mean responses of different groups we performed a multivariate ANOVA for each response mode, with firm size as a between subjects factor, scenario as a within subject factor, and dummy variables for various demographic groups as covariates. There were no significant differences between demographic groups, with the important exception of gender. Although the mean judgments of men and women are very highly correlated, women were somewhat more severe than men: they expressed higher outrage ratings (a mean difference of .52 scale units, $p < .001$), higher punishment ratings (difference of .37, $p < .001$), and set higher log damage awards ($p < .01$). There was also a gender X scenario interaction in which women assigned even higher ratings of outrage and punishment (but not higher dollar awards) to cases in which the plaintiff was female ($p < .05$).

In summary, we have evidence of a remarkable degree of consensus on the relative outrage evoked by scenarios of personal injury and on the relative severity of the punishment that is considered appropriate in this class of cases. This conclusion may not fully generalize to all domains of the law. We might expect to find much larger differences in the *level* of judged severity between communities and social categories in other areas, perhaps including attitudes to sexual harassment and civil rights violations, although there could still be agreement across categories on the *relative* severity of different instances within a domain.

3.1.4. From outrage to punishment: The harm effect. In the outrage model of Figure 1, morally objectionable actions evoke a graded response of outrage or indignation. An action can be judged more or less outrageous without reference to its consequences. Consequences, however, are important to punishment in law, and we suspected that they would also be important to lay intuitions about the proper punishment for reprehensible actions. This prediction was tested in Part 3 of the questionnaire, where different respondents encountered versions of the same scenario that differed in the harm that the plaintiff had suffered.

The results are summarized in the three panels of Figure 2. In each panel, the mean response to each high-harm scenario is plotted against the mean response to the low-harm version of the same scenario. Because outrage is assumed to be independent of consequences, our model predicts that the points should fall near the identity line. This prediction is confirmed. An analysis of variance shows no main effect of harm on outrage ratings. A different pattern is found for the punishment measure, which is displayed in the middle panel of Figure 2. Here the points all fall above the identity line, indicating that punishment ratings reflect the severity of the harm as well as the outrageousness of the action that caused it. The effect of harm is small, but statistically reliable ($F = 7.4$, $p < .01$). Because the outrage model involves a cascade of judgments, the effect of harm on

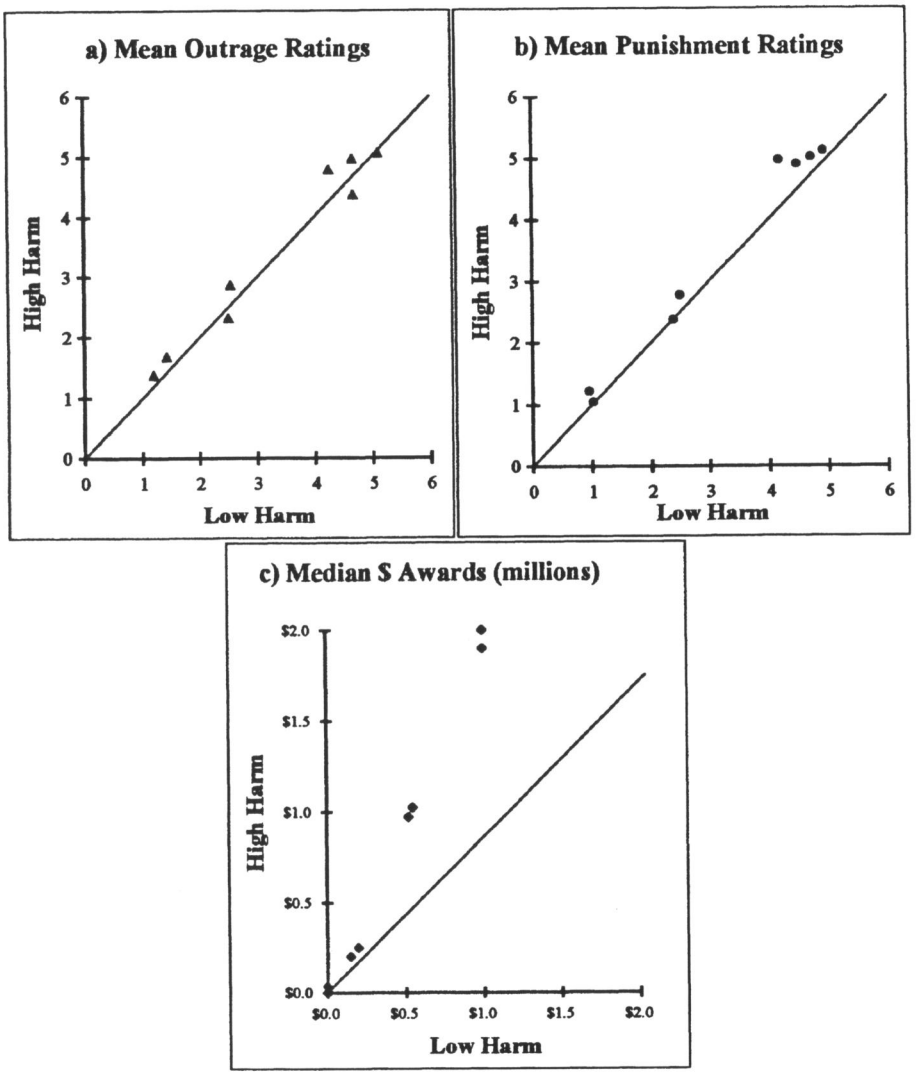

Figure 2. Comparison of Responses to Low and High Harm Versions of Same Scenario

punitive intent is expected to transfer to dollar awards as well, and indeed there is a statistically significant effect of harm on the natural log of damage awards ($F = 4.4$. $p < .05$). There is also a Harm X Firm Size interaction ($F = 6.6$, $p < .01$) in which the effect of Harm on log awards is enhanced for large firms.

3.1.5. From punitive intent to dollar awards: The effect of firm size. Within the academic community, opinion is sharply divided on the question whether the amount of punitive awards should depend on the size of the defendant firm (Polinsky and Shavell, 1997; Galanter and Luban, 1993). Lay intuitions, in contrast, are quite clear. In the outrage model, punitive intent is an intention to inflict pain; this means that the size of the defendant matters a good deal. Concepts such as "attracting the attention" of a corporation or inflicting pain on it are quite meaningful in everyday language, although philosophers, economists and legal scholars may find these concepts elusive. To ridinary people it seems intuitively obvious that it takes a louder message to attract the attention of a large firm and a larger loss to inflict pain on it. The outrage model therefore implies that the same intent to punish will be translated into a larger punitive award if the defendant firm is large than if it is small.

The four panels of Figure 3 present the relevant data. The logic of the display is the same as in Figure 2: the mean judgment for the large-firm version of each scenario is plotted against the mean judgment for the medium-firm version of the same scenario. The outrage model predicts that firm size should have no effect on ratings of outrage or intent to punish, or on the likelihood of deciding to make a non-zero award. Indeed, the data for all three measures fall symmetrically about and close to the identity line. In contrast, a large effect of firm size is apparent in the fourth panel. Analyses of variance yield a significant effect of firm size on log awards ($F = 4.8$, $p = .03$) and also on raw dollar awards ($F = 6.1$, $p = .01$).

The data reported in this section confirm the existence of broadly shared social norms about the acceptable behavior of firms (see also Kahneman, Knetsch and Thaler, 1986). They also lend strong support to the outrage model of punitive awards, which modulates a basic evaluation of a defendant's action by incorporating an evaluation of the harm that was caused to the plaintiff, and of the amount of loss that may be required to inflict an intended level of pain on a defendant.

The analyses presented so far were concerned entirely with the average judgments of the general public, and of sub-populations within it. At this level of aggregation we found consensus on social norms, and also concluded that the outrage model provides a rather precise account of moral judgments. However, there is considerable variability and skewness in individual judgments of dollar awards, and therefore also in the average judgments of small samples of people. Variability, especially the possibility of a very large award, is a critical problem in the world of legal practice, because juries are small samples. The next section is concerned with an analysis of variability.

3.2. The unpredictability of dollar awards

A central goal of the present study was to understand the causes of unpredictability in punitive damage awards. Our main hypothesis was that the mapping of punitive intent onto a dollar scale is a major cause of unpredictability, because of the highly skewed distribution of responses associated with magnitude scales. We expected to find much less variability and little skewness in the mapping of outrage and punitive intent onto category

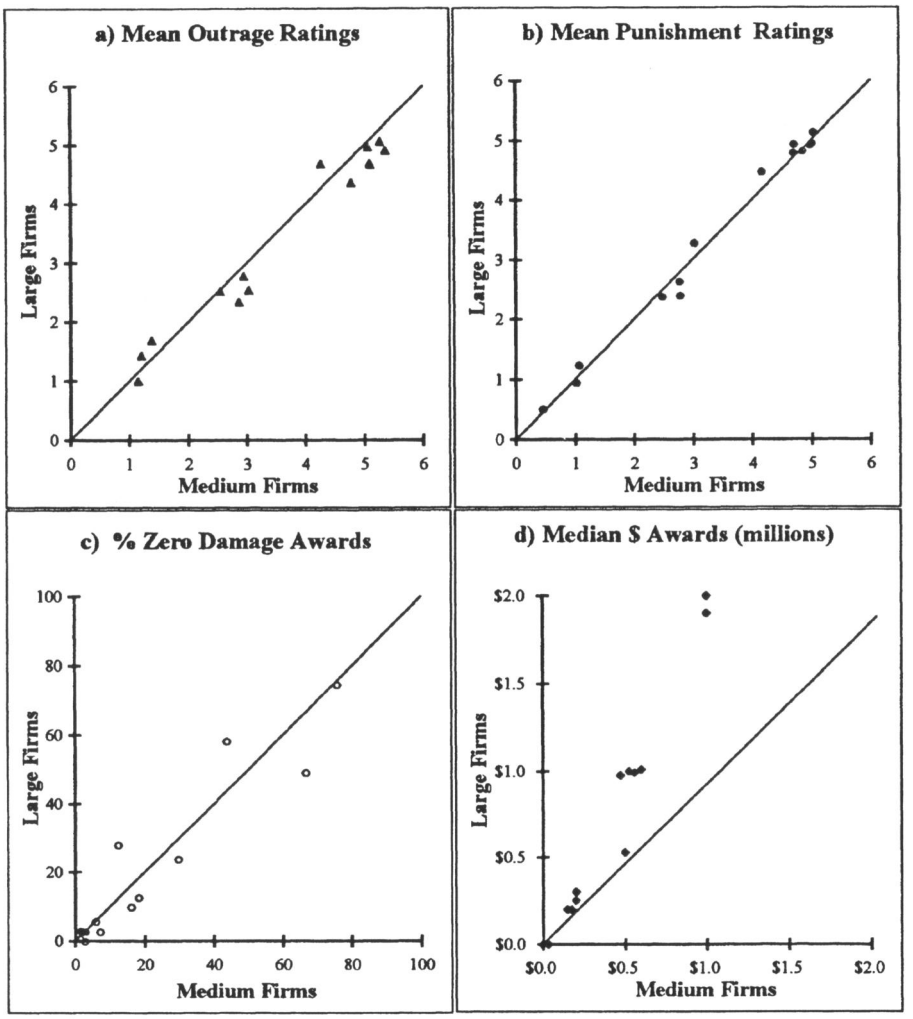

Figure 3. Comparison of Responses to Medium and Large Firm Versions of Same Scenario

scales, because these scales are bounded, and because descriptive labels guide respondents in the task of finding a response that expresses their intentions. In the following sections we present several analyses that compare these judgments.

3.2.1. Measuring unpredictability: Variance analysis. Our main concern is the extent to which the respondents in our study (and jurors in the real world) share the same evaluations of various scenarios involving tortious behavior by a defendant. In the ideal case, all respondents would agree in their evaluations of each scenario. Differences among sce-

Table 6. Proportion of Variance in Individual Responses Accounted for by Scenarios for Different Response Metrics[a]

Response Metric	Outrage	Punishment	$ Awards
Raw responses	.29	.49	.06
Convert to ranks for each subject	.42	.58	.51
Ln($)			.42

[a]Proportions were computed separately for each subgroup of subjects who evaluated exactly the same ten cases (including firm size and the harm levels in Part 3 scenarios). Table entries are means of the proportions from subgroups in the eight firm size (2) X harm condition (4) combinations. Subgroups ranged in size from 32 to 40 subjects.

narios would be the only source of variance in judgments—in this ideal situation there is only signal, no noise. In a non-ideal world, however, we expect to observe variance from two other sources: (1) individual differences in the average response to scenarios, and (2) "noise," which combines idiosyncratic responses of individuals to particular scenarios with pure measurement error. The proportion of total variance that is accounted for by scenarios provides a useful index of the extent to which judgments approximate the ideal of complete consensus. Table 6 presents the values of this index for the three responses used in the present study, and for some transformations of these responses.

To produce the values shown in Table 6, separate analyses of variance were conducted for each group of respondents who were exposed to the same combination of 10 of the 28 different scenarios. The values shown are a weighted average of the results obtained in these homogeneous groups. Ratings of outrage and punishment show much higher consensus than dollar awards. Indeed, only 6% of the variance of dollar awards is associated with scenarios. Similar results were reported by Kahneman and Ritov (1994), who compared willingness to pay for various causes (a magnitude sale) to category scale measures of related attitudes.

Dollar awards exhibit all the characteristics that are expected of magnitude scaling without a modulus, where each respondent in effect adopts a personal modulus. Individual differences in this modulus combined with the proportionality between individual means and standard deviations are expected to produce a high correlation between the mean and the standard deviations of respondents' judgments: if two individuals agree in their underlying impressions but their moduli differ by a factor, both the means and the standard deviations of their expressed judgments will differ by the same factor. For each group of respondents who saw the same set of 10 scenarios we computed the correlation between the mean and the standard deviation of the 10 responses made by each individual. As expected, this correlation was very high for dollar awards (an average of correlation of .90), but small for outrage ratings and (an average of .16) and punishment ratings (an average of .09).

Second, we observe that the distributions of awards are highly skewed to the right, for each of the 28 scenarios. We trace the skewness to two sources: the characteristic skewness of all magnitude judgments (Stevens, 1975), and the skewness in the distribution of personal moduli. We computed skewness statistics for non-zero awards for each scenario

for each group of respondents exposed to the same 10 scenarios, using the standard α_3 skewness measure.[2] Dollar award distributions were highly skewed to the right (an average skewness statistic of 3.57), while outrage ratings (an average of −.30) and punishment ratings (average of −.13) showed little skewness. To estimate the skewness of personal moduli, we computed the skewness index of the distribution of the means and of the standard deviations of positive individual responses, within each of the groups who saw the same set of scenarios. The average skewness was 4.00 for the means and 3.59 for the standard deviations. Finally, we generated a new set of judgments for each scenario within each group of respondents, by dividing all the positive responses of an individual by that individual's mean response. This transformation is an attempt to eliminate the effects of individual differences in moduli on the distribution of judgments. The average skewness of judgments within scenarios is now 1.35. This value is positive, reflecting the intrinsic skewness expected for magnitude judgments obtained with a specified modulus, but is much lower than the skewness of untransformed judgments. We conclude that the extreme skewness observed in the judgments of particular scenarios is mainly due to the skewed distribution of personal moduli.

In a further test of the hypothesis that the poor performance of the dollar measure is attributable to individual differences in moduli, we transformed each individual's 10 judgments to ranks. The ranking transformation eliminates individual differences in the use of the scale, while retaining the essential comparative information contained in the judgments. The transformation yielded a dramatic improvement in the predictability of dollar awards: the proportion of systematic variance rises from .06 to .51. In contrast, a ranking transformation yielded only a modest improvement in the proportion of scenario-related variance for outrage and punishment ratings. This pattern of results reinforces the conclusion that the unpredictability of raw dollar awards is produced primarily by large (and possibly meaningless) individual differences in the use of the dollar scale.

A logarithmic transformation of dollar awards also yields a dramatic improvement in the proportion of variance accounted for by scenarios. This result is to be expected for a magnitude scale. The transformation of awards to logs essentially eliminates the characteristic skewness of the distribution of judgments on these scales. Indeed, in their careful analysis of the predictability of actual punitive damage awards, Eisenberg et al (1997) showed that the logarithmic transformation yields a distribution that is almost normal. They also showed that the log of awards is predicted reasonably well from a set of objective characteristics of cases in which such awards were made. Eisenberg et al concluded that the unpredictability of punitive awards has been overstated. We find no inconsistency between their analyses of real jury awards and our experimental data. Indeed, we agree with their conclusion that log awards are reasonably predictable. Defendants and plaintiffs, however, live in a world of dollars, not log dollars. In terms of dollars the judgments of our respondents and of the real juries examined by Eisenberg et al are correctly described as erratic and unpredictable, because the high variability and severe skewness creates the possibility of disastrous losses which might induce risk aversion even in very large firms.

3.3. Synthetic jury analyses

The unpredictability that is of interest to the legal system concerns the award decision of a group of twelve (and sometimes fewer) citizens, who are given considerable information about a single case, and an opportunity to debate it at length. The circumstances of our experiment are of course quite different. The presence of multiple jurors provides the possibility of mitigating high individual variability to some extent through aggregation. In an attempt not to overstate the variability that is the object of our study, we chose to analyze the judgments of synthetic juries, derived from the responses of the individual respondents. The 'jury judgments' that we analyze are obtained by randomly selecting a group of 12 individuals, using the median response in the group as an estimate of the judgment that a jury composed of these 12 individuals would have rendered.

Our decision to use the median of individual judgments as an estimate is based on research findings which suggest that this statistic provides the best available estimate of the decision that is likely to emerge from group deliberations (Davis, 1996).[3] Diamond and Casper (1992) report an extensive examination of juror-jury relationships for compensatory damages in a price-fixing case. They also found that the median of the predeliberation responses of individual jurors was a better predictor of mock jury verdicts than either the mean or the mode.

There is no reason to believe that our main findings would be altered by the process of group deliberation. The conclusion of a now considerable literature using a wide variety of tasks is that deliberating groups hold no generalized advantage over individuals in the performance of judgment tasks (Kerr, MacCoun and Kramer, 1996). More specifically, Kerr, MacCoun and Kramer (1996) concluded from their review of judgmental biases in legal contexts that jury deliberations were actually slightly more likely to amplify the biases of individuals than to attenuate them. In the jury damage assessment context, Diamond and Casper's (1992) mock jury awards were significantly higher than the mean individual award (by 26%), a finding mirrored qualitatively in other studies (Davis, 1996, experiment 1; Kaplan and Miller, 1987, for punitive but not compensatory damages). If this effect applies to our study as well, the right-skewed distribution of damage awards in our data would cause our synthetic jury judgments to underestimate both the mean and the variance of the awards that that deliberating juries would have made. However, in a study with student mock-jurors Davis et al (1997) found that the median somewhat overestimated actual group judgments. While the relationship between individual judgments and that of the groups they compose is still an evolving area of research, this evidence suggests that the results we observe in our synthetic juries provide a reasonable and conservative approximation to the results that would be observed with real juries.

We created synthetic juries by randomly selecting groups of 12 respondents, with replacement, in each condition and for each scenario, and used the median of the 12 responses to represent the jury judgment (the size of the pool from which juries were drawn varied from 144 to 151 in Part 1–2 scenarios and from 64 to 77 in Part 3 scenarios). One thousand simulated juries were created in this fashion for each scenario and condition.

Table 7. Synthetic Jury Response Distributions by Scenario

Scenario	Firm Size	Harm Level	Lower 95% Confidence Bound	Median	Upper 95% Confidence Bound	Mean Jury Punishment	Prediction Error Ratio ($/Punish)
Joan	Large	High	$500,000	$2,000,000	$15,000,000	5.14	3.36
Joan	Medium	High	200,000	900,000	3,000,000	5.03	2.27
Thomas	Medium	—	200,000	500,000	1,575,000	5.02	1.69
Martin	Medium	High	350,000	1,000,000	4,000,000	4.98	4.01
Thomas	Large	—	200,000	560,000	2,750,000	4.95	.50
Joan	Large	Low	175,000	1,000,000	12,500,000	4.93	13.57
Martin	Large	High	350,000	1,900,000	10,000,000	4.92	2.40
Frank	Medium	—	230,000	760,000	2,100,000	4.86	1.67
Frank	Large	—	225,000	1,000,000	4,000,000	4.82	2.62
Mary	Large	—	290,000	1,000,000	4,000,000	4.79	1.49
Joan	Medium	Low	150,000	750,000	5,500,000	4.71	9.90
Mary	Medium	—	250,000	710,000	2,100,000	4.70	2.51
Martin	Large	Low	350,000	1,000,000	5,000,000	4.47	3.63
Martin	Medium	Low	200,000	675,000	2,250,000	4.16	2.53
Susan	Large	—	100,000	300,000	1,000,000	3.27	1.78
Susan	Medium	—	50,000	225,000	800,000	3.03	.93
Janet	Medium	High	100,000	200,000	690,000	2.79	1.37
Carl	Medium	—	15,000	155,000	375,000	2.78	1.59
Carl	Large	—	50,000	200,000	750,000	2.64	1.59
Janet	Medium	Low	0	150,000	650,000	2.49	2.00
Janet	Large	High	0	287,500	1,500,000	2.39	4.41
Janet	Large	Low	12,500	200,000	1,000,000	2.38	1.30
Jack	Large	High	0	0	350,000	1.24	2.10
Jack	Medium	High	0	45,000	225,000	1.07	1.30
Jack	Medium	Low	0	0	112,500	1.03	0.89
Jack	Large	Low	0	2,550	500,000	0.95	2.91
Sarah	Large	—	0	0	1,000	0.51	1.12
Sarah	Medium	—	0	0	13,000	0.46	∞
						Median	2.18

Table 7 presents, for each of the scenarios used in the study, the median and the lower and upper bounds of a 95% confidence interval for the awards of simulated dollar juries, as well as the mean ratings of simulated punishment juries. The first observation is that the unpredictability and characteristic skewness of jury dollar awards is readily replicated under laboratory conditions.[4] Note that the extreme variability of awards is especially striking for the most severe cases. As noted earlier, this result could be anticipated from similar observations with magnitude scales in the context of psychophysics, where the standard deviation of judgments is commonly found to be proportional to the mean (Stevens, 1975). In the data of Table 8, the width of the confidence interval is highly correlated with both the mean (r = .90) and the median (r = .87) of dollar awards.

3.3.1. Measuring unpredictability: correlational analysis. Jury judgments can be considered predictable, in our use of that term, if there is high agreement between juries randomly selected from the population. We carried out an analysis that compares the

Table 8. Median Correlations between Judgments of Simulated Juries

	Outrage	Punishment	$ Awards
Outrage	.87		
Punishment	.86	.89	
$ Awards	.47	.51	.42
Overall Median $ Award	.71	.77	.69

predictability of the judgments made by simulated dollar juries, outrage juries and punishment juries. The basic unit of this analysis is a set of 28 jury judgments, one each for each of the 28 scenarios listed in Table 7. Except for the possibility of the same individuals appearing in more than one jury, this procedure simulates the convening of independent juries to deal with 28 separate cases on the same day. Using this procedure, we created 60 sets of 28 simulated jury judgments for each response mode, where each judgment was contributed by a different jury. We then computed the Pearson correlation between each pair of sets. This computation was performed both within response (e.g., the correlation between the outrage ratings of successive juries) and across responses (e.g., the correlation between the outrage rating of one jury and the punishment rating of another). The data shown in Table 8 are medians of the 1770 correlations obtained within each response mode or of the 3600 correlations obtained between two response modes.

The results of Table 8 conform to our hypothesis. There is strong agreement between independent synthetic juries judging outrage or punishment on category scales (r = .87 and .89). In contrast, agreement between independent synthetic juries judging dollar awards on a magnitude scale is quite weak (r = .42). Evidently, the individual variability and skewness of judgments on this measure is so large than even the medians of 12 judgments are unstable. The problem could be reduced, of course, by taking larger samples. For example, we found that when the size of the juries is increased to 30, the correlation between the dollar awards of independent sets of juries rises to .80. However, the problem would be exacerbated in smaller samples, such as juries of size 6 or 8.

A jury can be thought of as a sample from the community, whose function is to provide an estimate of community sentiment. In the context of our experiment, community sentiment about the punitive damages for a scenario will be defined as the median of the damages set by all the respondents who judged it. The bottom row of Table 8 presents the correlations between sets of simulated jury judgments for the 28 scenarios and corresponding estimates of community sentiment, for which we used the overall median of dollar awards for each scenario. Remarkably, punishment juries predict community sentiment about dollar awards more accurately than dollar juries do. This superior accuracy is especially notable because, as will be shown below, the relationship between dollar awards and ratings of punitive intent is highly non-linear.

3.3.2. Improving predictability. As noted above, the judgments of dollar juries can be viewed as estimates of community sentiment about the punitive award that is appropriate in a given case. The accuracy with which a dollar jury estimates community sentiment can be assessed by computing the difference between the jury's award and a measure of

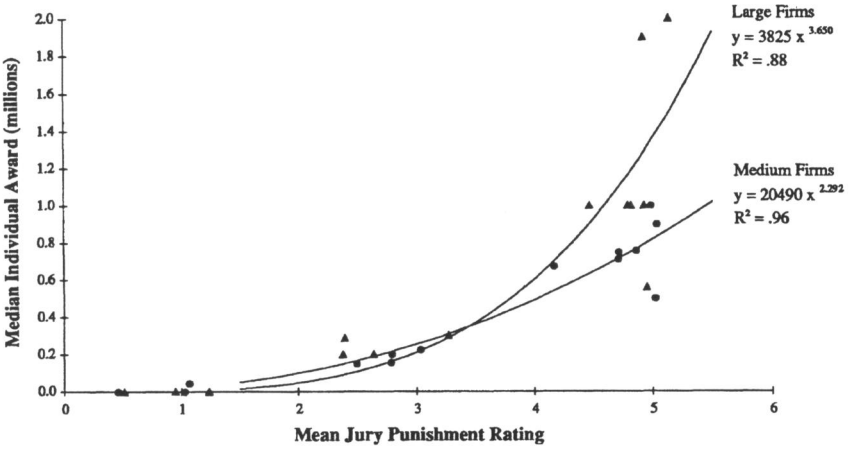

Figure 4. Mapping Intended Punishment to Community Damage Awards for Synthetic Juries

community sentiment, for which we use the overall median of individual dollar awards. The root-mean-square of these discrepancies (RMSE) across juries, calculated separately for each scenario, provides a measure of average prediction error that is analogous to the standard error of the estimate in a regression.

Hypothesis 5 states, and the correlational analysis summarized in the preceding section implies, that punishment juries can be used to obtain more accurate estimates of community sentiment about dollar awards than dollar juries themselves can provide. To examine this possibility, we proceed as follows. Using the entire sample, we first estimate the functions that relate the overall median of dollar awards for a given scenario to the mean judgment of all simulated punishment juries for that scenario (see Figure 4). Next, we create punishment juries according to the re-sampling procedure described above, and use the median punishment rating of the 12 individuals as the simulated jury judgment. We then use the relevant function from Figure 4 to convert the judgment of each punishment jury into a dollar award. As before, the computed awards are treated as estimates of the overall median of dollar awards, and we use RMSE as an index of the accuracy of these estimates.

As can be seen in Figure 4, the relationship between punishment ratings and dollar awards is highly non-linear. On the background of previous psychophysical scaling research, this result is not surprising. Stevens (1975) pointed out that magnitude scales are generally related to category scales of the same underlying variable by a power function. Following Stevens, we therefore fitted (separately for medium and large firms) conversion functions of the following form:

$$D_i = aP_i^b + e \tag{1}$$

where D_i = median jury award for scenario i, and P_i = mean jury punishment for scenario i. Mean punishment ratings lower than 1.5 were interpreted as intending no punishment, and a prediction of $0 was made for these scenarios. Conversion functions were estimated on those scenarios that exceeded this threshold, and produced good fits (R^2 = .96 and .88 for medium and large firms, respectively). We specified a squared error loss function and estimated a and b using a nonlinear numerical search routine. Figure 4 presents the data and the estimated functions for this analysis.

To assess the relative quality of the estimates from jury dollar awards and from converted punishment ratings, we created 100 juries of 12 individuals for each response mode and each scenario, in the same fashion as before (with replacement) and computed an estimate of the population median of dollar awards. For dollar juries, this estimate was simply the median of the twelve individual responses. For punishment juries, the median of individual responses was converted to an estimated dollar award using the conversion function for the appropriate firm size. The squared differences between these estimates and the observed median dollar award for each scenario were summed across the 100 juries. Finally, we computed for each scenario the ratio of the RMSE for dollar awards to the RMSE for converted punishment ratings. The last column of Table 7 presents these ratios, which we use as a summary statistic of the relative accuracy of the two estimation procedures.

As is apparent from the large ratios (the median ratio is 2.18), the estimates derived from punishment ratings are far more accurate than estimates from dollar awards. For example, for the case of Joan with a medium firm size and high harm, the ratio is 2.27 (Table 7), the median individual award is $1,000,000 (Table 4), and the estimates of dollar juries have an average error from this value of $913,481 compared to $402,414 for estimates based on punishment juries. A difference in this direction was found for 25 of the 28 scenarios. These results support the rather counter-intuitive conclusion that the goal of estimating community sentiment about dollar awards can be achieved more efficiently by asking juries to rate their intent to punish than by asking them to set dollar awards—provided of course that the requisite conversion function is available.

4. Discussion

The research we have reported was designed with three objectives in mind: (i) to learn about the psychology of punitive damages; (ii) to understand why punitive damages are sometimes arbitrary and unpredictable; (iii) to consider possible reforms in the task assigned to the jury, which might reduce this uncertainty. We discuss the three objectives in turn.

4.1. The psychology of damage awards

We have presented a model in which judgments about punitive awards are interpreted as a succession of mappings. The core of the model is an emotion of outrage, which we have

made no attempt to analyze in this research—we were content with measuring it. This elementary internal state is a central determinant of the responses to many other questions that could be asked about the same incident: for example the emotion that is expressed in a rating of appropriate punishment or in setting dollar awards would also affect evaluations of the image of a defendant firm and of the character of individual defendants.

Although many questions about the case would evoke the same emotion, each question also raises a particular set of considerations. The outrage model is a particular type of attitude structure in which several internal states or responses are organized in a chain of increasing complexity; a more complex response depends on all the factors that affected its predecessor in the chain, and also on some new factors. In this structure the emotion of outrage is basic, punitive intent is intermediate and overt judgments of dollar awards are most complex. The determinants that affect any link in this chain are a superset of the determinants of earlier links.

We have identified some of the determinants of punitive intent and of dollar awards, but the list we offered is not complete. For example, the harm suffered by plaintiff is surely not the only factor that influences punitive intent beyond outrage. In particular, the designation of an action as a repeated offense could well produce a more severe punishment even if it does not increase the outrageousness of the act. Similarly, firm size is surely not the only factor that affects dollar awards beyond punitive intent. Numerical anchors are another likely candidate. Because there is no obvious way to determine what it might take to punish a company "severely" or "very severely", the mapping from punishment to dollars is likely to be highly susceptible to suggestions from other people and to arbitrary anchors (Chapman and Bornstein, 1996; Hastie, Schkade and Payne, 1997). Indeed, we cannot reject the interpretation that the firm size effect that we observed in the present study was produced by anchoring on the numbers that were provided to characterize firms of different sizes. More generally, we view the outrage model not as a finished product but as a framework for future investigation of the intuitions that jurors bring to their task.

The normative status of the outrage model raises difficult questions. On the one hand, this model appears to capture moral norms that are widely shared in society. On the other hand, there are reasons to question whether the social norms that produce outrage deserve to be honored by the legal system in awarding punitive damages. The backward-looking, retributive focus of outrage is not well-suited to the deterrence function that many legal theorists view as the main justification for punitive damages (Landes and Posner, 1993; Polinsky and Shavell, 1997). Those who emphasize deterrence generally urge that punitive damages should be awarded to make up for the likelihood that some acts will not be detected and punished. This idea is not represented in the outrage model. Indeed, we suspect that the brazen actions of an agent that commits reprehensible acts in the full light of day may occasion more outrage than the actions of a more surreptitious perpetrator. Thus, efforts to make punitive awards serve deterrent goals will run up against the intuitions that jurors bring to their role. Legitimate questions can also be raised about the intuitions that treat corporations as persons, and large corporations as big and powerful persons. In some areas, legislatures committed to retribution may want to question the social norms of lay persons. We return to these issues below, where we raise the possibility of a radical change in the role of juries in setting punitive awards.

4.2. A cause of unpredictable awards

The task of setting punitive damages most closely resembles a particular psychophysical method, which we have labeled **magnitude scaling without a modulus.** As was noted earlier, subjects in most applications of magnitude scaling are instructed to assign a particular number (the modulus) to a specified standard stimulus, and to judge other stimuli in terms of ratios of subjective intensity relative to that standard. The modulus is both arbitrary and inconsequential: the same ratio scales will be obtained regardless of whether the standard stimulus is to be judged 10 or 100. However, magnitude scaling can also be conducted without a modulus, by instructing subjects to ensure that their judgments of different stimuli should reflect the ratios of the intensity of subjective experience. In the context of psychophysical scaling (e.g., of loudness), the procedure yields scales that are quite similar to those that are obtained in the standard method, though somewhat noisier and more skewed. It appears that a subject in such an experiment spontaneously adopts a modulus—and different subjects adopt different moduli. Naturally, this procedure greatly increases the variability and skewness of the numerical responses to any given stimulus. The effects of individual differences in the size of the modulus are easily eliminated in the statistical analysis of psychophysical data, where each observer judges many stimuli. Under current legal practice, this remedy is not available, since each jury considers only one case in isolation, and indeed is explicitly prohibited from comparing it to others.

The close similarity of our results to the pattern observed in psychophysical tasks raises two questions: (i) what is the significance of individual differences in the overall magnitude of responses? (ii) what is the status of the average of the judgments assigned to a given scenario (or stimulus)? In the context of psychophysics, the answers to these questions are (i) individual differences in the size of responses reflect differences in the choice of an arbitrary modulus; and (ii) the average responses to any single stimulus are also largely arbitrary, although the ratios of responses to different stimuli are meaningful.

Do these conclusions also apply to the setting of punitive damages in our study? Was the mapping from punitive intent to dollars as labile and arbitrary in our experiment as it would be in a psychophysical experiment? In the present study, individual differences in the overall magnitude of awards need not be completely arbitrary: they could reflect genuine differences of opinion about the punitive effects of punitive damages, that is, in the amount required to engage the attention of negligent firms or to inflict pain on them. The data of the present study do not permit us to estimate how much of the variability of dollar awards is due to such systematic effects. However, there are several reasons to believe that arbitrary differences in individual choices of modulus play a major role. The first is the evidence indicating the great sensitivity of punitive awards to anchors of dubious informative value. In addition, we have seen that the mapping of punitive intent onto dollars raises normative and conceptual problems that present a challenge to the expert. It is likely that lay persons faced with these problems will make rather arbitrary choices. The normative relevance of dollar damage awards is limited by the extent to which the moduli used by jurors are arbitrary.

We have considerable confidence in the conclusion that arbitrary differences in the use of the dollar scale were an important cause of the unpredictability of dollar juries in our study. A more difficult question concerns the possibility of extrapolating from the artificial circumstances of the present experiment to the real legal world. The critical fact, we propose, is that actual juries and the participants in our experiments share the task of mapping punitive intent onto an unbounded scale, with no specified standards to guide them. Under these circumstances we expect judgments to be highly labile, and therefore susceptible to any anchors that may be provided in the course of the trial or in jury deliberations.

A final note of caution is in order here. The fact that punitive damages share the known deficiencies of magnitude scaling is likely to be a significant cause of unpredictable punitive awards—but it is not the only one. Other relevant factors include regional differences, plaintiff's demand, the quality of the lawyers on both sides, and doubtless others.

4.3. Implications for legal reform

In spite of the acknowledged limitations of our experiment, we believe that our findings have substantial implications for reforms of the role of juries in setting punitive awards. Of course the decision of whether and how the current system should be reformed must be guided by a normative analysis of the social and legal function of punitive damages. In the subsequent discussion we distinguish among three canonical positions, which differ in the diagnosis of the principal weakness of the current system, and imply different recommendations for how it might be improved. The issues raised in the following sections are discussed in greater detail in Sunstein, Kahneman and Schkade (1998). In that paper, we also discuss other reform proposals, such as damage caps, compensatory damage multipliers, and an increased supervisory role for judges.

4.3.1. High variability of dollar awards. The first and least radical of the three positions we shall discuss assumes that the only problem with the punitive awards made by juries is that they are erratic and unpredictable, because of the large individual differences between jurors in the mapping of punitive intent into dollars. In this view the "community sentiment" about the appropriate dollar punishment in a case is the normative standard. The task of the jury is simply to reflect this sentiment, which we have operationalized here as the median of the judgments that would be made if the entire eligible population participated in setting dollar awards. The problem of unpredictability arises from an identifiable source: a sample of twelve is too small to provide an accurate estimate of community sentiment, because of the large individual differences in the mapping of punitive intent. The goal of reform is merely to obtain a more accurate understanding of that sentiment.

How might this problem be solved? The analysis that we described in Section 3.3.2 showed that asking juries to assess severity of punishment, and converting this assessment to dollars by a conversion function leads to much greater accuracy (by a factor of more than 2 in these data) than asking juries to set dollar awards. The procedure of this

experiment provides a blueprint for a manageable program of research and reform that could yield a substantial improvement in the predictability of jury awards.

In the first phase of such reform, a set of fictitious but realistic scenarios would be generated—perhaps several sets corresponding to different types of case, such as personal injury or financial misconduct. These cases would be presented to a large number of individuals (or experimental juries) to obtain estimates of population norms. Some of the respondents would set dollar awards. Others would provide a rating of punitive intent, or perhaps rank the scenarios by this variable.[5] The results of this study would provide a conversion function that could subsequently be used in real cases to transform jury ratings of punishment (or their ranking of the case at hand in a standard set of calibrated cases) into dollar awards. The analysis that we reported can be seen as a partial simulation of this new procedure. As we have seen, the results of this simulation suggests that error variance can be reduced by changing the jury's task to set punishment in more abstract terms (by rating or by ranking), and avoiding the less manageable task of setting dollar awards. Specifically, we found that it is possible to exploit the error pattern of category scales (where the error variability and the mean are not generally correlated) to achieve a more accurate estimate of community sentiment than the current system of asking juries to set dollar awards can provide.

The proposed procedure raises many questions of implementation. From the normative point of view, however, it is straightforward because it accepts the principle that punitive damages should reflect the central tendency of community sentiment. This is a quite modest critique of the existing system, and as a reform proposal it represents a modest change, providing improved estimates of community judgments about dollar amounts without questioning the normative force of those judgments.

If we pursue this perspective, then other means of improving the quality of our estimate of community sentiment can be considered as well. For example, one possibility suggested to us by Dan McFadden involves adding the observed mean damage award of the relevant category (e.g., personal injury cases) to the jury judgment as a second predictor of the community median. This approach can further reduce error variance by capitalizing on the stability of the category mean. Also, we did not study other possible measures of punitive intent and it is possible that further investigation could identify an even more stable and reliable scale, which would further improve the prediction of community sentiment about dollar awards. Many new possibilities are opened by raising the question "How can we obtain the best estimate of community sentiment?"

4.3.2. Inadequate calibration of dollar awards. A more radical critique would argue that the legal system should deny the normative status of community sentiment about dollar awards, but accept community sentiment about punitive intent. If the translation of punitive intent onto the dollar scale is arbitrary and poorly informed, the goal of the jury system should not be to predict community sentiment about dollars—even if this prediction could be made accurate.

On this view, the conversion functions that take punitive intent to dollars (see Figure 4) reflect intuitions about appropriate dollar awards that are neither informed nor particularly reasonable. Even if it is granted that the jury is sovereign is deciding how much an

individual or a corporation should be made to "suffer", what knowledge do jurors have that would enable them to convert their punitive intent into a particular dollar number? Even experts would find that question difficult to answer. As we have seen, it is quite plausible that individuals and juries answer it by generating arbitrary moduli, or by clinging to numerical anchors provided in the information they receive. If community sentiment on awards is seen as representing the averaged effects of arbitrary individual moduli, another procedure will be required to translate punitive intent into dollar awards.

The proposal that naturally emerges from this analysis again involves a two-phase determination of punitive awards. As before, juries should be asked to provide a rating or a ranking of the case at hand on a scale of punitive intent, as in the preceding proposal. However, the conversion function from punishment to dollars would be generated by a legislative or regulatory process, not by the empirical mapping from one measure of community sentiment to another. The conversion function would take into consideration the factors that affect community judgment, such as firm size, but it would be based on an informed analysis of the harm that different levels of damage would cause. The availability of a formal scheme to convert punitive intent into dollars would not only reduce the unpredictability of awards; it is also likely to make them more fair and more reasonable.

4.3.3. Rejecting the outrage model. A still more radical critique of the current system could draw on our findings about the moral intuitions that are embodied in the outrage model. Our results suggest that these intuitions have a backward-looking, retributive focus; that they appear to neglect issues of deterrence; that they treat corporations as persons; and that they require punitive damages to be sensitive to the financial status of the defendant. From the normative standpoint, some people have questioned each of these aspects of the outrage model. Many observers believe that the social function of punitive damages should be deterrence, not retribution (Landes and Posner, 1993; Polinsky and Shavell, 1997). The view of the corporation as a person is problematic, and perhaps less than perfectly robust: would potential jurors be swayed by the information that the people who ultimately pay the punitive damages are stockholders, employees, and customers, most of them individually powerless to change the behavior of the corporation? Questions have also been raised about the justification for the common intuition that punitive damages should be adjusted to the size of the defendant.

It appears that jurors bring to their task some strongly ingrained and broadly shared moral intuitions, which may not correspond to the actual or appropriate legal principles. Accepting this position not only denies the legitimacy of dollar awards; it also denies the legitimacy of the punitive intent of the community. Could jury instructions point the jury in the right direction? Perhaps; but a large literature on the difficulty of changing attitudes suggests that the intuitions of the outrage model are likely to affect jurors' thinking, even in the face of explicit instructions to the contrary (see Hastie, Schkade and Payne, forthcoming).

If accepted, the logic of this analysis would lead to a proposal that appears radical in the special context of American legal tradition and practice: Juries should not be assigned the task of setting punitive damages, because they are unlikely to carry out their task appropriately in light of the best understanding of the goal of punitive awards. It follows that the

task of determining the size of awards should be turned over to judges or to some administrative process. If a jury is used at all, its function would be to make a judgment about the outrageousness of the defendant's behavior, which would be used as one of several inputs in a punitive decision that would be made by some other person or agency (see Sunstein, Kahneman and Schkade, 1998).

5. Concluding remarks

We have shown that the difficulty of mapping punitive intent onto dollar amounts is a potentially important source of unpredictability in punitive damages. The variability of the dollar response contrasts with another important observation: there is substantial agreement in the population with respect to judgments of both outrage and punishment, and the consensus operates across differences of gender, race, age, education, and income. This substantial consensus on outrage and punitive intent make them predictable, while dollar awards are much less so.

This finding leads naturally to some specific reform proposals. In particular, it suggests that any effort to improve predictability in the award of punitive damages will be most successful if it assists with the task of mapping outrage and punitive intent onto a dollar scale. A conversion formula might well be adopted to provide this assistance. The content of the conversion formula depends on the normative status that is accorded to community sentiments. As we have seen, different recommendations for reform emerge from three different diagnoses of the problem: (i) dollar awards are susceptible to excessive sampling error; (ii) the mapping of punitive intent to dollars is poorly informed; (iii) the judgments of jurors are likely to be affected by outrage and by retributive urges that are not in line with appropriately designed legal principles.

Our findings have implications well beyond the area of punitive damages, because the legal system is pervasively in the business of requiring people to map normative judgments onto dollar amounts. Our analysis suggests that the mapping of attitudes onto unbounded scales creates problems that are both predictable and severe. We suspect, for instance, that similar difficulties exist in the setting of awards for pain and suffering, libel, and civil rights violations, and in attempts to use measures of willingness to pay to estimate the value of public goods.[6] There is a need to develop mechanisms for eliciting social preferences and values that do not suffer from the problems and defects identified here. We do not expect to find an uncontroversial way of eliciting, or constructing, social preferences and values, but some approaches to this important task are demonstrably worse than others, because they make people perform tasks for which they are ill-equipped. The problems we have identified and the solutions we have sketched here are likely to be relevant to many questions currently facing law and policy.

6. Appendix I: Excerpts from instructions

In this study we would like you to imagine that you are a juror for a legal case in a civil court. Civil law suits can involve disputes between private individuals, companies, or individuals and companies, in which the plaintiff alleges that the defendant harmed them or their property in some way. A civil suit is brought by a plaintiff for the purpose of gaining compensation from the defendant for the alleged harm.

Civil suits involve two different types of penalties that could be imposed upon a defendant that is found liable for damages. *Compensatory damages* are intended to fully compensate a plaintiff for the harm suffered as a result of the defendant's actions. *Punitive damages* are intended to achieve two purposes: (1) to *punish* the defendant for unusual misconduct, and (2) to *deter* the defendant and others from committing the same actions in the future.

In the cases you will consider, the defendant has already been ordered to pay compensatory damages to the plaintiff. This does not necessarily mean that punitive damages must also be awarded. Whether or not punitive damages should be awarded and, if so, how large they should be, is completely separate from compensatory damages.

Punitive damages should be awarded if a preponderance of the evidence shows that the defendants either acted maliciously or with reckless disregard for the welfare of others. Defendants are considered to have acted *maliciously* if they intended to injure or harm someone. Defendants are considered to have acted with *reckless disregard* for the welfare of others if they were aware of the probable harm to others but disregarded it, and their actions were a gross deviation from the standard of care that a normal person would use.

Civil suits differ from criminal cases, in which the government prosecutes an individual or a company for alleged violations of the law. Plaintiffs in a civil trial must prove their claim by "a preponderance of the evidence," which means that it is more likely than not that the plaintiff's claim is justified. This differs from criminal trials, where the prosecution must prove the defendant's guilt "beyond a reasonable doubt."

In each of the cases you will consider, the jury (of which you are a member) has already decided to accept the plaintiff's claim. As a consequence the jury has ordered the defendant to pay $200,000 in compensatory damages to the plaintiff as full compensation. The defendants are large [medium-sized] companies with profits of $100–200 [$10–20] million per year.

7. Appendix II: Personal injury scenarios

Mary Lawson

Mary Lawson, a manufacturing worker, developed chronic anemia. Although after a hospital stay she is now better, the condition has not been fully cured. She believes that exposure to benzene in her work place caused the condition and sued her employer, TGI

International. The jury (of which you are a member) ordered TGI International to pay her $200,000 in compensation.

TGI International is a large company (with profits of $100–200 million per year) that manufactures high-tech machine parts. Some years ago the scientists at TGI International discovered that manufacturing workers at Mary Lawson's plant were often exposed to benzene, a substance that can cause anemia, leukemia and cancer. Internal documents show that the top management at TGI International decided not to do anything about the problem, because benzene levels in the plant were slightly below the maximum level allowed by OSHA regulations. They thought that the risk was worth taking and that "with any luck no one will get sick." They also decided against warning the workers, because "warnings would just create panic."

Frank Williams

Frank Williams suffered serious internal injuries when the braking system on his motorcycle failed to work at a traffic light. He felt that that the brakes were defective, and sued National Motors, the company that manufactures and sells his motorcycle. The jury (of which you are a member) ordered National Motors to pay him $200,000 in compensation.

National Motors is a large company (with profits of $100–200 million per year) that makes motorcycles, scooters, and other motorized single person vehicles. The evidence showed that the braking system used by National Motors has a basic design defect. National Motors was aware that "there might be a problem with our brakes," because in pre-market tests, the defect appeared on several occasions. But the pre-market tests were not extensive, despite the fact that auto industry regulations require elaborate testing. Internal company documents show National Motors' belief that "it would be quite expensive for us to do much more now, we can't be certain we have a serious problem here, and anyway we can fix the problem afterwards if it really does turn out to be serious."

Thomas Smith

While he was visiting the circus, Thomas Smith was shot in the arm by a security guard who mistakenly thought that Smith had threatened another customer with bodily harm. The security guard was drunk at the time. Smith sued Public Entertainment, the company that operates the circus. The jury (of which you are a member) ordered Public Entertainment to pay him $200,000 in compensation.

Public Entertainment is a large company (with profits of $100–200 million per year) which operates circuses and public fairs. Fred Williams, the security guard who was involved in the incident, is an alcoholic with a history of incidents of drunkenness on the job. During one of these incidents Williams took out his gun and started waving it around wildly, but he did not shoot anyone. Public Entertainment had repeatedly warned Williams

to "clean up his act" but took no other action. In his company personnel file Williams was described as "basically a good guy with a bit of a drinking problem, but not enough of a risk to fire him."

Susan Douglas

Susan Douglas suffered significant injuries to her legs and neck when an airbag in her car opened unexpectedly while she was driving the vehicle. She believes that the airbag was defective and sued the manufacturer, Coastal Industries. The jury (of which you are a member) ordered Coastal Industries to pay her $200,000 in compensation.

Coastal Industries is a large company (with profits of $100–200 million per year) that specializes in parts and accessories that can be added to existing vehicles, such as adding the latest safety equipment to older cars. While its airbag conforms to the requirements stated in government regulations, it does not include certain additional "fail-safe" systems that are used in other airbags to ensure against accidental opening. Internal documents show that most but not all of the Coastal Industries designers believed that their system "is certainly safe enough, even if it does not include all possible safeguards" and that their marketing department said that "there will be no market for our airbag if we raise its price by adding more safety bells and whistles."

Carl Sanders

Carl Sanders used Nalene, an over-the-counter baldness treatment available at drugstores. While a small amount of hair did grow back, he also developed severe side-effects, including open sores on the scalp and permanent brown spots over his forehead. He sued the manufacturer, A&G Cosmetics. The jury (of which you are a member) ordered A&G Cosmetics to pay him $200,000 in compensation.

A&G Cosmetics is a large company (with profits of $100–200 million per year) that sells many different cosmetic products, including wigs, "weaves" and chemical solutions designed to combat baldness. Nalene has proven effective in promoting hair growth in 30% of people in clinical trials. However, Nalene caused unpleasant side effects in some cases, although none were as severe as those Carl Sanders experienced. When marketing Nalene, A&G Cosmetics did not fully disclose these findings. It only said "minor side effects have been observed in a very small number of people tested." While this amount of disclosure was within legal limits, other companies that make hair products voluntarily disclosed more about their products.

Sarah Stanley

Sarah Stanley, a seventy-five year old woman, suffered serious back injuries as a result of following an exercise video, "Good Health For All," that she purchased through her local

community health center. When Stanley attempted to perform the exercises, she found herself unable to do so, but she pressed on beyond her physical capacities. She claimed that she was not adequately warned of these dangers and sued the producer of the video, Gersten Productions. The jury (of which you are a member) ordered Gersten Productions to pay her $200,000 in compensation.

Gersten Productions is a large company (with profits of $100–200 million per year) that produces informational materials in health-related fields, including videos on many topics concerning healthy lifestyles. The "Good Health for All" video contains a series of exercises suitable mostly for people in good shape and good health. The exercise coaches and models in the video are all relatively young, and no federal or state law requires exercise videos to come with any special warning for elderly people. The witnesses in the case testified that Gersten Productions believed that most people would be able to tell when the exercises were beyond their capacities, that Good Health for All has produced good results for almost all people who have seen it, and that very few people had reported injuries of any type from doing so.

Jack Newton

Jack Newton, a five year old child, was playing with matches when his cotton flannel pajamas caught fire. He was severely burned over a significant portion of his body and required several weeks in the hospital and months of physical therapy. His parents sued the manufacturer of the pajamas, Novel Clothing. The jury (of which you are a member) ordered Novel Clothing to pay the Newtons $200,000 in compensation.

Novel Clothing is a large company (profits of $100–200 million per year) that specializes in making clothes for children. Before marketing the pajamas, Novel conducted the tests normally used in the industry for problems like flammability, and observed no incidents like the Newtons experienced. Companies in the industry as well as federal regulators have known for a while that it is possible to add extra fire-retardant chemicals to their fabrics (in addition to those specified in current regulations), but these extra measures are not required. The process is very expensive, and no other manufacturers currently use it. Internal documents show that the management of Novel Clothing had decided that "when it comes to costly safety innovations we will follow our competitors. We don't want to be less safe than anyone else but we don't have to lead the way either."

Low harm version: Jack Newton, a five year old child, was playing with matches when his cotton flannel pajamas caught fire. His hands and arms were badly burned, and required regular professional medical treatment for several weeks.

Joan Glover

Joan Glover, a six year old child, ingested a large number of pills of Allerfree, a non-prescription allergy medicine, and required an extensive hospital stay. The overdose weak-

ened her respiratory system, which will make her more susceptible to breathing-related diseases such as asthma and emphysema for the rest of her life. The Allerfree bottle used an inadequately designed child-proof safety cap. The Glovers sued the manufacturer of Allerfree, the General Assistance company. The jury (of which you are a member) ordered General Assistance to pay the Glovers $200,000 in compensation.

General Assistance is a large company (with profits of $100–200 million per year) that manufactures a variety of non-prescription medicines. A federal regulation requires "child-proof" safety caps on all medicine bottles. General Assistance has systematically ignored the intent of this regulation by selling tens of thousands of bottles of medicines with a child-proof safety cap that was generally effective, but had a failure rate much higher than any others in the industry. An internal company document says that "this stupid, unnecessary federal regulation is a waste of our money;" it acknowledges the risk that Allerfree may be punished for violating the regulation but says, "the federal government has many other things to worry about and probably won't bother us on this" and in any case "the punishments for violating the regulation 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 vice president of General Assistance that they were "on shaky ground" but the company decided not to take any corrective action.

Low harm version: Joan Glover, a six year old child, ingested a large number of pills of Allerfree, a non-prescription allergy medicine. She had to spend several days in a hospital, and is now deeply traumatized by pills of any kind. When her parents try to get her to take even beneficial medications such as vitamins, aspirin, or cold remedies, she cries uncontrollably and says that she is afraid.

Martin West

Martin West, a right-handed disabled veteran who lived in a two story house, was seriously injured in a fall when the chain broke on his electric lift-chair (a device that allows someone to be carried up stairs in a chair that moves up and down an angled track). He fell from near the top of the stairs and tumbled awkwardly all the way to the bottom landing, damaging his spinal cord in the process. As a result he now has only partial control of his right arm, a condition which doctors believe is permanent. He sued the manufacturer of the lift-chair, MedTech Products. The jury (of which you are a member) ordered MedTech to pay him $200,000 in compensation.

MedTech Products is a large company (profits of $100-200 million per year) that manufactures many types of medical equipment, including wheelchairs, car-lifts, and other devices used by the disabled. The lift-chair is a new product for MedTech, and instead of producing a new design, company engineers decided to adapt the design of the hydraulic lifts for cars already on the market. Unfortunately, there are several unique problems in designing a safe and effective lift-chair that are beyond the experience of the company's engineers. Product managers said that hiring new engineers with the proper expertise was "too expensive, and would take too long" and ordered current engineers to

"just do the best you can, but be sure you meet our deadline for announcing the product." The inexperience of the engineers and the rush to meet the product announcement date led to testing procedures that were less rigorous than those required by federal medical product regulations.

Low harm version: Martin West, a left-handed disabled veteran who lived in a two story house, was injured in a fall when the chain broke on his electric lift-chair (a device that allows someone to be carried up stairs in a chair that moves up and down an angled track). He fell from near the bottom of the stairs and tumbled to the bottom landing, injuring his spinal cord in the process. His right arm was paralyzed for several weeks, after which doctors were able to repair most of the injury, and he was able to regain most of the previous range of motion in the arm.

Janet Windsor

Janet Windsor, a secretary who works on computer equipment, developed a rare form of skin cancer. After a long course of painful chemotherapy, doctors were able to cure the cancer, although they cannot be sure that it will not return. She believed that it had been caused by the computer monitors that she worked on and sued the manufacturer, International Computers. The jury (of which you are a member) ordered International Computers to pay her $200,000 in compensation.

International Computers is a large company (profits of $100–200 million per year) that manufactures components of computer systems. The type of International Computers monitor that Ms. Windsor used emits an unusually high level of radiation compared to other similar monitors, a level that pushes the limit in government safety guidelines. Internal company documents cite experts who concluded that "the evidence that this level of radiation could create any serious risk to health and life is weak and tentative." The company was not legally required to disclose the unusual level of radiation, and it did not do so.

Low harm version: Janet Windsor, a secretary who works on computer equipment, suffered from frequent and severe migraine headaches. As a result, for several years she often experienced nausea, insomnia and depression, and missed many workdays and family events.

Acknowledgment

This research was supported by Exxon Company, U.S.A. The data reported and the opinions expressed in this article belong to the authors. John Payne, Lawrence Mark, Lawrence Ward, Dan McFadden and Rob MacCoun provided helpful comments on an earlier draft. We also thank Ron Carrell for his able assistance with the administration of the study.

Notes

1. Except for cases at the extremes of the scale, where all respondents agree.
2. α_3 is based on the ratio of the third to the second moments of the distribution, and is obtained by calculating a z-score (i.e., substract the sample mean and divide by the sample standard deviation) for each observation, cubing it, and then computing the average of these cubed z-scores. A positive value means the distribution is skewed to the right, a negative statistic means it is skewed to the left, and zero means it is symmetric.
3. Davis and his colleagues have developed a model called the Social Judgment Scheme (SJS) that may hold promise for improving on the median, but there is as yet no clear evidence for this claim. For example, in Davis (1996) the SJS was slightly better than the median at predicting the responses of individual juries but slightly worse at predicing the overall distribution of awards.
4. On the extent to which actual punitive damage awards are unpredictable, compare Eisenberg (1997) with Polinsky (1997).
5. The observation that a context of comparable cases improves the signal/noise ratio of judgments suggests that ranking could be more useful than rating; see section 3.1.1.
6. The design and some of the analysis of the present study were patterned after an earlier study of willingness to pay for public goods (Kahneman and Ritov, 1994).

References

Bargh, John. A., Shelley Chaiken, Paula Raymond, Charles Hymes. (1996). "The Automatic Evaluation Effect: Unconditional Automatic Attitude Activation with a Pronunciation Task," *Journal of Experimental Social Psychology* 32, 104–128.

Bell, David, Howard Raiffa, and Amos Tversky (Eds). (1988). *Decision Making: Descriptive, Normative and Prescriptive Interactions*. Cambridge, UK: Cambridge University Press.

Chapman, Gretchen, and Brian Bornstein. (1996). "The More You Ask for the More You Get: Anchoring in Personal Injury Verdicts," *Applied Cognitive Psychology* 10, 519–540.

Davis, James H. (1996). "Group Decision Making and Quantitative Judgments: A Consensus Model." In E. Witte and J. Davis (Eds.), *Understanding Group Behavior: Consensual Action by Small Groups*. Mahwah, N.J.: Erlbaum.

Davis, James H., Wing Tung Au, Lorne Hulbert, Xiao-ping Chen, and Paul Zarnoth. (1997). "Effects of Group Size and Procedural Influence on Consensual Judgments of Quantity: The Example of Damage Awards and Mock Juries," *Journal of Personality and Social Psychology* 73, 703–718.

Diamond, Shari, and Jonathan Casper. (1992). "Blindfolding the Jury to Verdict Consequences: Damages, Experts, and the Civil Jury," *Law and Society Review* 26, 513–563.

Eisenberg, Theodore, John Goerdt, Brian Ostrom, David Rottman, and Martin Wells. (1997). "The Predictability of Punitive Damages," *Journal of Legal Studies* 26, 623–662.

Galanter, Marc, and David Luban. (1993). "Poetic Justice," *American University Law Review* 42, 1393–1453.

Hampton, Jean. (1993). "The Retributive Idea." In J. Hampton and J. Murphy (Eds.), *Forgiveness and Mercy*. Cambridge University Press: Cambridge.

Hastie, Reid, David Schkade, and John Payne. (forthcoming). "A Study of Juror and Jury Judgments in Civil Cases: Deciding Liability for Punitive Damages," *Law and Human Behavior*.

Hastie, Reid, David Schkade, and John Payne. (1997). "Effects of Plaintiff Identity and Plaintiff's Damage Request on Juror Assessments of Punitive Damages," Working paper.

Huber, Peter. (1989). "No-fault Punishment," *Alabama Law Review* 40, 1037–1049.

Jeffries, John. (1986). "A Comment on the Constitutionality of Punitive Damages," *Virginia Law Review* 72, 139–151.

Kahan, Daniel, and Martha Nussbaum. (1996). "Two Conceptions of Emotions in Criminal Law," *Columbia Law Review* 96, 269–374.

Kahneman, Daniel, and Ilana Ritov. (1994). "Determinants of Stated Willingness to Pay for Public Goods: A Study in the Headline Method," *Journal of Risk and Uncertainty* 9, 5–38.

Kahneman, Daniel, and Ilana Ritov. (1998). "Preferences, Attitudes and Dollars," *Journal of Risk and Uncertainty*, in press.

Kahneman, Daniel, Ilana Ritov, Karen Jacowitz, and P. Grant. (1993). "Stated Willingness To Pay for Public Goods: A Psychological Analysis," *Psychological Science* 4, 310–315.

Kahneman, Daniel, and Jack Knetsch. (1992). "Valuing public goods: The purchase of moral satisfaction," *Journal of Environmental Economics and Management* 22, 57–70.

Kahneman, Daniel, Jack Knetsch, and Richard Thaler. (1986). "Fairness as a constraint on profit seeking: Entitlements in the market," *The American Economic Review* 76, 728–741.

Kahneman, Daniel, Jack Knetsch, and Richard Thaler. (1990). "An Experimental Test of the Coase Theorem," *Journal of Political Economy* 98, 1325–48.

Kaplan, Martin, and Charles Miller. (1987). "Group Decision Making and Normative Versus Informational Influence: Effects of Type of Issue and Assigned Decision Rule," *Journal of Personality and Social Psychology* 53, 306–313.

Kerr, Norbert, Robert MacCoun, and Geoffrey Kramer. (1996). "Bias in Judgment: Comparing Individuals and Groups," *Psychological Review* 103, 687–719.

Landes, William, and Richard Posner. (1993). *Economic Analysis of Tort Law*. Cambridge: Harvard University Press.

Lodge, Milton. (1981). "Magnitude Scaling: Quantitative Measurement of Opinions," in J. Sullivan (Ed.), *Quantitative Applications in the Social Sciences* 25, Sage Publications. Beverly Hills.

Polinsky, A. Mitchell, and Steve Shavell. (1997). "Punitive Damages: An Economic Analysis," *Harvard Law Review*.

Stevens, Stanley S. (1975). *Psychophysics. Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley: NY.

Sunstein, Cass, Daniel Kahneman, and David Schkade. (1998). "Assessing Punitive Damages," *Yale Law Journal*, May.