Article

Concordance and Conflict in Intuitions of Justice

Paul H. Robinson† and Robert Kurzban††

Introduction 1830
I. Are Lay Intuitions of Justice Specific? 1832
A. Categorization Studies 1837
B. Ranking Studies 1838
C. Magnitude Estimation Studies 1839
D. Counterarguments 1840
E. Small Factual Changes Producing Significant Differences in Liability Judgments 1842
II. Is There Agreement Among People's Intuitions of Justice? 1846
A. The Intuition that Serious Wrongdoing Should Be Punished 1848
1. Questionnaire Studies 1848
2. Behavioral Economics Studies 1850
3. Cross-Cultural Studies 1852
4. Are There Exceptions? 1853
B. Intuitions on the Relative Seriousness of Wrongdoing 1854
1. Previous Domestic Studies 1855
2. Cross-Cultural Studies 1862
III. Testing the Limits of Agreement 1866
A. Extent of Agreement: Study 1 1867
1. Method 1867
2. Results 1868

† Colin S. Diver Professor of Law, University of Pennsylvania. The authors are indebted to Sarah Robinson for data collection, to Michael Orchowski, Alex Shaw, Emma Chen, Peter DeScioli, and Lindsay Suttenberg for invaluable research assistance, and to Daniel Houser for assistance in data analysis.

†† Assistant Professor of Psychology, University of Pennsylvania. Copyright © 2007 by Paul H. Robinson and Robert Kurzban.

3. Discussion 1872
B. Extent of Agreement: Study 2 1874
1. Method 1874
2. Results 1876
3. Discussion 1880
IV. Disagreements on Intuitions of Justice 1880
A. Apparent Disagreements Among Intuitions of Justice 1880
B. True Disagreements Among Intuitions of Justice: Studies 3 and 4 1883
1. Method: Studies 3 and 4 1883
2. Study 3 Results 1884
3. Study 4 Results 1887
4. Discussion 1890
Conclusion 1892
Appendix A 1894
Appendix B 1900
Appendix C 1902
Appendix D 1904

INTRODUCTION

The role of justice in assigning criminal liability and punishment has been a mainstay of philosophical, legal, and policy debate for centuries. Most of the past century's debate has used a philosophical concept of justice, drawn from the reasoned analysis of moral philosophers. Some more recent writings, however, have urged that there might be good reason to rely upon a more empirical notion of justice: one reflected in the shared intuitions of justice of the community to be governed by the criminal justice system whose rules and practices are being formulated.[1]

1. *See, e.g.*, PAUL H. ROBINSON & JOHN M. DARLEY, JUSTICE, LIABILITY, AND BLAME: COMMUNITY VIEWS AND THE CRIMINAL LAW 6–7 (1995) [hereinafter ROBINSON & DARLEY, JUSTICE, LIABILITY, AND BLAME] ("Greatest cooperation will be elicited when the criminal law's liability rules correspond with the community's views of justice."); Paul H. Robinson & John M. Darley, *The Utility of Desert*, 91 NW. U. L. REV. 453, 457–58 (1997) (arguing that the power of the criminal justice system to influence conduct is rooted in its moral credibility, and that deviations from community norms of desert detracts from this credibility by creating perceptions that the system is unfair). For a discussion of the difference between philosophical and empirical desert, see Paul H. Robinson, Competing Conceptions of Modern Desert: Vengeful, Deontological, and Empirical (Feb. 8, 2007) (unpublished manuscript, on file with the authors) [hereinafter Robinson, Competing Conceptions]; Paul H. Robinson, *The Role of Moral Philosophers in the Competition Between Deontological and Empirical*

But many writers, including many sympathetic to the importance of doing justice—that is, of distributing punishment according to a robust assessment of an offender's blameworthiness—see desert as a vague concept on which there is much disagreement.[2] One line of writers articulates a related but slightly different objection: It is not that desert is too vague to have any meaning or that people necessarily disagree about what desert means. Rather, they argue, the notion of desert has discernible content but content that provides guidance only as to the outer limits of appropriate punishment. It can identify injustice and failures of justice—outer boundaries that should not be crossed—but cannot specify a punishment that *should* be imposed.[3]

Because, until recently, the traditional defense of a desert distribution of punishment has been philosophical, there has been little effective response to such attacks regarding specificity, vagueness, and disagreement. Indeed, the disagreements among the desert-supporting philosophers themselves only seem to illustrate the validity of the criticisms.[4] But when an empirical conception of desert is adopted, the claims of vagueness and disagreement are matters subject to empirical testing. Do lay persons have specific intuitions of justice, specific enough to reliably distinguish among a variety of cases? Or, does the empirical conception of desert have hopelessly vague content or does it specify only matters of outer limits of proper punishment, such that it cannot be the basis for constructing a workable criminal justice system? Even if the empirical notion of desert has meaningful and specific content, is that meaning so different among different persons that, again, it could not be used to construct a workable criminal justice system? In other words, are there *shared* intuitions of justice?

Desert (Univ. of Pa. Law Sch. Pub. Law, Working Paper No. 06-39 (2007)), *available at* http://ssrn.com/abstract=933692.

2. *See infra* notes 5–6 and accompanying text.

3. *See infra* notes 5–6 and accompanying text.

4. *See, e.g.*, MARK TUNICK, PUNISHMENT: THEORY AND PRACTICE 107 (1992):

> From our account of theorists commonly called retributivists it is clear that there is no distinct set of criteria the satisfaction of which is essential for meriting that label. . . .
>
> The retributivist label, then, might not seem particularly useful, for the differences on particular issues among some retributivists may seem greater than the differences between some retributivists and some utilitarians.

Id.

Parts I and II of this Article seek to answer these questions by examining previously existing research done in this and other countries. Part III reports the results of a new set of studies that attempted to find the limits of agreement concerning the core wrongs with which criminal law concerns itself. Finally, Part IV examines areas in which there are true disagreements among people's intuitions of justice. It also examines the reasons why the level of agreement that does exist may be obscured from the view of the casual observer, creating a false impression of disagreement.

The picture that emerges is quite striking: Intuitions of justice among laypersons exist on a wide variety of liability and punishment issues. They are quite nuanced, no matter a person's level of education. They produce specific directions regarding deserved punishment, not simply broad generalities or outer limits. And there is a good deal of agreement on intuitions of justice regarding a wide range of liability and punishment issues and across all major demographics.

I. ARE LAY INTUITIONS OF JUSTICE SPECIFIC?

Writers commonly criticize what they see as the "vagueness of desert."[5]

> [E]veryone may agree that five years in prison is unjustly harsh desert for shoplifting, or that a five dollar fine is unjustly lenient desert for rape, but beyond such clear cases our intuitions seem to fail us. Is two years, five years, or ten years the proper sanction for a rape? . . . Our sense of just deserts here seems to desert us.[6]

5. *See, e.g.*, BERNARD BOSANQUET, SOME SUGGESTIONS IN ETHICS 188, 203 (1918) ("There is no estimate which can determine degrees of moral guilt in actual individual cases. Such a thing is wholly inconceivable." (emphasis omitted)); JOHN BRAITHWAITE & PHILIP PETIT, NOT JUST DESERTS: A REPUBLICAN THEORY OF CRIMINAL JUSTICE 179 (1990) ("The vagueness of desert . . . masks mistakes."); R.A. Duff, *Penal Communications: Recent Work in the Philosophy of Punishment*, 20 CRIME & JUST. 1, 8 (1996) ("It is not enough simply to appeal to the supposedly shared intuition that the guilty deserve to suffer . . . since such an intuition, however widely shared, needs explanation: *what* do they deserve to suffer, and *why*?" (citation omitted)).

6. Leo Katz, *Criminal Law*, *in* A COMPANION TO PHILOSOPHY OF LAW AND LEGAL THEORY 80, 80–81 (Dennis Patterson ed., 1996). Similarly, it is observed:

> Perhaps, at best, retributivism can determine the roughly appropriate punishment by comparatively ranking offenses in such a way as murder warrants greater punishment than rape, which warrants greater punishment than armed robbery, and so on. But it cannot determine whether rape warrants twenty, thirty, forty years imprisonment. Though retributivism cannot set cardinal or absolute levels of pun-

Further, it is argued, even if theoretically there was a meaningful concept of desert, it simply would be impossible to operationalize it with sufficient specificity:

> Insofar as we seek a morally sensitive scale in which to weigh subjective guilt, to classify the individual criminal on the long continuum from unblemished virtue to unmitigated evil . . . . [t]he criminal law is unfitted for such issues. It faces an adequacy of difficulties without addressing such ethical nuances. It is necessarily generalized rather than related to the moral quality of the specific act. . . .
>
> Questions of guilt will thus be weighed on the imprecise scales of the criminal law which can allow for only a few subjective qualifications to the objective gravity of the crime.[7]

> ishment, its advocates insist that they can set ordinal, or relative, levels of punishment (for example, murder warrants greater punishment than larceny). But retributivism cannot even satisfactorily determine degrees of punishment ordinally. For example, even if we assume that, all other things being equal, murder warrants greater punishment than armed robbery, does negligent homicide warrant greater punishment than intentional rape or intentional armed robbery? Retributivism has no answer. This is the serious flaw in Kant's lex talionis and even G.W.F. Hegel's more sophisticated version. Both fail to take into account differing culpability levels stemming from the various levels of mens rea or mental states that accompany the commission of wrongdoing. Retributivism has no answer to the issue of whether greater wrongdoing done with lower culpability (for example, negligence or recklessness) warrants more or less punishment than comparatively minor wrongdoing with a greater level of culpability (such as intention or purpose). Thus, retributivism can determine neither the ordinal nor the cardinal ranking of crimes and their concomitant degrees of punishment.

Russell L. Christopher, *Deterring Retributivism: The Injustice of "Just" Punishment*, 96 NW. U. L. REV. 843, 893 (2002) (footnotes omitted).

7. NORVAL MORRIS, THE FUTURE OF IMPRISONMENT 74 (1974). Other writers have expressed similar views in different terms:

> Norms are necessarily stated in general terms, in the sense of applying to classes of circumstances, defined fairly generally. Accordingly, there are many opportunities for disagreement to arise among members of a society over whether a norm is applicable in a specific circumstance. For example, it is likely that virtually all Americans hold that rape is prohibited behavior, but whether a specific instance of sexual intercourse is definable as rape may be unclear because the boundary between acceptable and unacceptable means of persuasion is not fixed. Indeed, a large part of the work of our legal system consists of making determinations of whether and how legal norms apply to specific instances of behavior. Accordingly, by the very nature of norms, we can expect greater consensus about them when they are stated in general terms and lesser consensus over the application of the norms to specific instances of behavior. In the present instance, we can expect more agreement among Americans about which crimes deserve harsher punishments but less agreement about the specific sentences to be imposed.

PETER H. ROSSI & RICHARD A. BERK, JUST PUNISHMENTS: FEDERAL GUIDELINES AND PUBLIC VIEWS COMPARED 2–3 (1997).

Some writers may be willing to concede that desert has some meaning, but argue that it cannot specify a particular amount of punishment that is deserved; it can only identify what is a seriously disproportionate punishment. Indeed, this is the underlying assumption of the American Law Institute's recent proposal for a revised Model Penal Code Section 1.02(2)(a), setting out the purposes of the sentencing provisions and the principles governing their interpretation and application:[8]

> Subsection (2)(a) embraces Morris's observation that moral intuitions about proportionate penalties in specific cases are almost always rough and approximate—and that most people experience them as such. Even when a decisionmaker is acquainted with the circumstances of a particular crime, and has a rich understanding of the offender, it is seldom possible, outside of extreme cases, for the decisionmaker to say that the deserved penalty is *precisely x*. In Morris's phrase, the "moral calipers" possessed by human beings are not sufficiently fine-tuned to reach exact judgments of condign punishments. Morris postulated instead that most people's moral sensibilities, concerning most crimes, will orient them toward a range of permissible sanctions that are "not undeserved." Outside the perimeters of the range, some punishments will appear clearly excessive to do justice, and some will appear clearly too lenient—but there will nearly always be a substantial gray area between the two extremes.

> Although there may be little difficulty in making uniform judgments of ordinal culpability (e.g., his killing was more culpable than her killing, because she was provoked and he wasn't) or of ordinal harmfulness (e.g., her theft was more harmful than his, because she stole their car and he stole their typewriter), there is no unique nonarbitrary way to combine these judgements into one judgement of ordinal seriousness.

Hugo Adam Bedau, *Retribution and the Theory of Punishment*, 75 J. PHIL. 601, 613 (1978).

8. The text of the proposal provides:

§ 1.02(2). Purposes; Principles of Construction.

(2) The general purposes of the provisions on sentencing are:

(a) in decisions affecting the sentencing of individual offenders:

(i) to render sentences within a range of severity proportionate to the gravity of offenses, the harms done to crime victims, and the blameworthiness of offenders;

(ii) in appropriate cases, to achieve offender rehabilitation, general deterrence, incapacitation, and restoration of crime victims and communities, provided these goals are pursued within the boundaries of sentence severity permitted in subsection (a)(i); and

(iii) to render sentences no more severe than necessary to achieve the applicable purposes in subsections (a)(i) and (ii) . . .

MODEL PENAL CODE § 1.02 (Sentencing Discussion Draft No. 1, 2006).

> Subsection 1.02(2)(a)(i) codifies Morris's conception of an approximate retributive ballpark when it speaks of a "*range* of severity" of proportionate punishments. Subsection 2(a)(ii) makes further reference to the idea of a permissive range when it refers to "the *boundaries* of sentence severity permitted in subsection (a)(i)."[9]

Part of the confusion here is a mistaken notion by some that desert associates some absolute amount of punishment with a particular offense. Modern desert theorists make no such claim and have no need to do so. They argue that desert requires that punishment be proportional to an offender's personal blameworthiness.[10] That is, desert does require a specific amount of punishment, not because there is some magical connection between that offense and that amount of punishment, but rather because that amount of punishment is the amount needed to set the offender in his appropriate relative position on the continuum of deserved punishment. In other words, modern notions of desert are ordinal rather than cardinal. If the continuum of punishment is altered—for example, some societies permit the death penalty and others do not, some use long prison terms while others do not—then desert would distribute punishment differently. Desert demands only that each person's offense be punished according to the person's relative degree of personal blameworthiness. There is nothing in desert theory that demands a particular, higher or lower, end point on the continuum of punishment. Desert primarily concerns itself with the relative difference in blameworthiness among cases.[11]

But it may still be argued that even this blameworthiness ranking of offenses is beyond the ability of people's intuitions of justice, that those intuitions are simply too vague to do more than to roughly distinguish between "serious" cases and "less

9. *Id.* § 1.02(2) cmt. b; *see also* Richard S. Frase, *Sentencing Principles in Theory and Practice*, 22 CRIME & JUST. 363, 365–70 (1997) (summarizing Norval Morris's limiting retributivist theory of punishment in which desert dictates upper and lower limits of punishment, but general deterrence, considerations of equality, and parsimony provide "fine-tuning").

10. JOHN KLEINIG, PUNISHMENT AND DESERT 114 (1973) ("What follows from [attempts to exactly define desert] is not the absurdity of specific desert claims but only the absurdity of expecting them to function like statements of empirical quantity."); ANDREW VON HIRSCH, PAST OR FUTURE CRIMES: DESERVEDNESS AND DANGEROUSNESS IN THE SENTENCING OF CRIMINALS 39–46 (1985) ("Desert should be treated as a determining principle in deciding *ordinal* magnitudes.").

11. A popular attack on desert is to ignore this character of modern desert and to insist that it demands a particular amount of punishment and, further, that it demands harsh punishment. Robinson, Competing Conceptions, *supra* note 1 (manuscript at 11–13, 18–21).

serious" cases and cannot provide the kind of nuance needed to construct a workable criminal justice system. As Justice Stephen Breyer explains in his opposition to basing the United States Sentencing Commission Sentencing Guidelines on desert:

> The "just deserts" approach would require that the Commission list criminal behaviors in rank order of severity and then apply similarly ranked punishments proportionately. For example, if theft is considered a more serious or harmful crime than pollution, then the thief should be punished more severely than the polluter. . . .
>
> [T]he Commission soon realized that only a crude ranking of behavior in terms of just deserts, based on objective and practical criteria, could be developed. . . .
>
> [T]hose who advocated "just deserts" . . . could not produce a convincing, objective way to rank criminal behavior in detail . . . .[12]

12. Stephen Breyer, *The Federal Sentencing Guidelines and the Key Compromises upon Which They Rest*, 17 HOFSTRA L. REV. 1, 15–17 (1988) (footnotes omitted). For other writers arguing that crimes cannot be rank ordered, see, for example, David Dolinko, *Three Mistakes of Retributivism*, 39 UCLA L. REV. 1623, 1638–39 (1992). Commissioner Paul H. Robinson argued, in response, that the Commission's compromise approach left the guidelines with no coherent underlying principle and, not surprisingly, internally incoherent in directions. *See* Paul H. Robinson, *Dissent from the United States Sentencing Commission's Proposed Guidelines*, 77 J. CRIM. L. & CRIMINOLOGY 1112, 1113–16 (1986):

> Neither of the Commission's guidelines was drafted with a coherent, articulated sentencing philosophy in mind. Rather, the drafting was done in an ad hoc manner without the guidance of any set of sentencing principles. The inevitable result of this approach is guidelines that are haphazard and internally inconsistent, and that frequently generate improper results; they simply do not consistently and rationally distinguish cases according to relevant offense and offender characteristics.
>
> A comparison of possible guideline sentences for different offenses illustrates one difficulty. Is it appropriate that the sentence for aggravated fish smuggling can be greater than that for armed bank robbery? that the sentence for aggravated forcible sexual contact with a 13-year-old child can be less that [sic] that for submitting a false record on protected wildlife? that the sentence for some antitrust violations can be less than that for failure to surrender a naturalization certificate? that the sentence for involuntary manslaughter can be less than that for impersonating a government employee? that the sentence for inciting a riot can be less than that for altering a motor vehicle ID number? The fact of the matter is that the Commission never systematically ranked offenses. . . .
>
> While the proposed draft obviously does not generate peculiar results in all cases, it is all too easy to find difficulties like those noted above. The true significance of these examples is not the particular problems that they present, but rather that they manifest an unsystematic approach to the complex task of guideline drafting. It is an approach that has produced a flawed structure and drafting of mixed quality.

The average state criminal code distinguishes a dozen grades of offenses.[13] Modern sentencing guidelines make even more distinctions.[14] Are lay intuitions of justice nuanced enough to distinguish among cases with this level of specificity or more? Or, do they suggest only broad categories?[15]

A. CATEGORIZATION STUDIES

A common measure of whether a person makes nuanced judgments is the number of distinctions she makes in her judgments. In studies in which subjects are given categories into which to sort offenses, they in fact use most or all of the categories. Thorsten Sellin and Marvin Wolfgang, for example, asked raters to evaluate the seriousness of violations using a scale ranging from 1 (least serious) to 11 (most serious), with instructions that "[e]ach of the eleven categories is an equal step on the scale."[16] Monica Walker surveyed 650 individual residents of Sheffield, England and forty first-year sociology

---

Id.

13. *See, e.g.,* ARIZ. REV. STAT. ANN. § 13-601 (2001) (distinguishing six categories of felonies, three categories of misdemeanors, and one category of petty offense); COLO. REV. STAT. § 18-1-104 (2006) (listing six felony categories, three misdemeanor categories, and two petty offense categories); KAN. STAT. ANN. §§ 21-4704 to -4708 (1995) (showing ten felony categories, divided into "drug" and "nondrug" offenses, and three misdemeanor categories); NEB. REV. STAT. §§ 28-105 to -106 (1995) (listing eight felony categories and seven misdemeanor categories).

14. The U.S. Sentencing Guidelines represent the far end of this spectrum, with forty-three offense levels. U.S. SENTENCING GUIDELINES MANUAL § 5A (2006), *available at* http://www.ussc.gov/2006guid/gl2006.pdf.

15. *See* MORRIS, *supra* note 7, at 31–33. The psychological literature on "categorical perception" provides an example of how Morris' categorization-only intuitions might work. Consider the perception of phonemes, small units of linguistic sounds such as "ba" and "pa." Careful experimentation has shown that what distinguishes these sounds in spoken language is "voice onset time," a continuous variable—a variable that can take on an arbitrarily large number of values. Despite this fact, these sounds are perceived as either one phoneme or the other. In intermediate cases, listeners do not report that the phoneme is somewhere between the two, but simply is one or the other. This illustrates that there are domains in which judgments are categorical rather than nuanced. *See* Alvin M. Liberman et al., *The Discrimination of Speech Sounds Within and Across Phoneme Boundaries*, 54 J. EXPERIMENTAL PSYCHOL. 358, 358–68 (1957) (studying the discrimination of speech sounds within and across phoneme boundaries by having subjects listen to two syllables and indicate if the two were the same or different; syllables were identified as different more easily if they were on opposite sides of a phoneme boundary).

16. THORSTEN SELLIN & MARVIN E. WOLFGANG, THE MEASUREMENT OF DELINQUENCY 131, 140 (1967).

students, asking subjects to rank the seriousness of eleven offense scenarios—such as "He steals £1 of the money and alters the books"—on a scale ranging from 1 to 11.[17] Stephen Gottfredson, Kathy Young, and William Laufer similarly used eleven categories of "seriousness" in their survey of 159 subjects.[18] In all three studies, subjects used the full range or nearly the full range of the scale values that were available to them.[19]

B. RANKING STUDIES

The more nuanced people's intuitions of justice, the easier it should be for people to see two cases as distinguishable. A variety of studies, including those reported here, asked participants to compare two or more cases or to rank cases against one another. We know of no instances in which subjects were presented with such a task and were unable to complete it. Walker, noted above, used the "paired comparison" method, in which subjects are asked to judge which in each pair of offenses

17. Monica A. Walker, *Measuring the Seriousness of Crimes*, 18 BRIT. J. CRIMINOLOGY 348, 348–49 (1978) (comparing relative rankings of crimes across social status and gender using category and ratio scales; showing that very serious and very trivial offenses were excluded because other studies had already shown substantial agreement between people with regard to these crimes; and also presenting evidence that the entire scale was used, even when the ranges of offenses were restricted—giving further evidence of nuanced intuitions).

18. Stephen D. Gottfredson et al., *Additivity and Interactions in Offense Seriousness Scales*, 17 J. RES. IN CRIME & DELINQ. 26, 29 (1980).

19. SELLIN & WOLFGANG, *supra* note 16, at 254; Gottfredson et al., *supra* note 18, at 29; Walker, *supra* note 17, at 349. In some cases, it is not possible to determine precisely which categories were used because the data were reported as means rather than entire distributions. This brief review is not meant to imply that this limits the number of cases in which offenses were assigned to categories of seriousness. Other examples include a study by Don Gibbons, who had people evaluate twenty offenses and used categories of punishment ranging from "no punishment" to "execution." All categories were used by at least some subjects. Don C. Gibbons, *Crime and Punishment: A Study in Social Attitudes*, 47 SOC. FORCES 391, 395 (1969) (detailing that only four of the twenty offenses yielded more than ten percent "no penalty" responses: homosexuality, consensual sex between a twenty-year-old and a sixteen-year-old, draft evasion, and marijuana use). Likewise, Peter Rossi used a similar method and obtained similar results. The 140 offenses rated fell across essentially the entire nine-category scale. Peter H. Rossi et al., *The Seriousness of Crimes: Normative Structure and Individual Differences*, 39 AM. SOC. REV. 224, 230 (1974). In our search of the literature we found no cases in which only a narrow band of the available categories was used by subjects. This search includes the studies reported in Parts I, II, and VI.

is more serious.[20] There is no evidence that participants had any difficulty with this task.[21]

C. MAGNITUDE ESTIMATION STUDIES

Another testing methodology—magnitude estimation—suggests the same conclusion. In Sellin and Wolfgang's survey of 575 individuals[22] across Pennsylvania, subjects were asked to rate the seriousness of fifty-one offenses and to rate the seriousness of each offense relative to a bicycle theft, which was given an arbitrary value of 10.[23] Thus, if an offense was judged twice as serious as a bicycle theft, that offense would be given a value of 20. Subjects' responses covered an extremely broad range of values covering nearly three orders of magnitude.[24]

In the 1977 National Crime Survey,[25] which surveyed sixty thousand persons across the United States, subjects were read vignettes of twenty-five specific criminal events[26] and asked to rate how serious the crime was relative to a baseline offense (the same bicycle theft used by Sellin and Wolfgang). Not all subjects rated all items; a total of 204 offenses were evaluated.[27] Judgments of the severity of the crimes covered an extremely broad range: two and a half orders of magnitude.[28]

---

20. Walker, *supra* note 17, at 355.

21. *See id.* at 361.

22. The study included 251 students, 286 police officers, and 38 judges. SELLIN & WOLFGANG, *supra* note 16, at 255–58.

23. Each subject rated the same twenty-one offenses and an additional thirty offenses drawn from a pool of 120 additional offenses. *Id.* at 255–57; *see also id.* at 268 ("The most strongly supported conclusion . . . is that all the raters . . . tended to . . . assign . . . estimations [so] that the seriousness of the crimes is evaluated in a similar way, without significant differences, by all the groups" and, further, that a "pervasive social agreement about what is serious and what is not appears to emerge. . . .").

24. Each order of magnitude refers to an offense being ten times more serious than another. Two orders of magnitude would mean that the most serious offense was considered one hundred times as serious as the least serious offense, and so on. *See id.* app. E-2 at 389–90.

25. *See* MARVIN E. WOLFGANG ET AL., THE NATIONAL SURVEY OF CRIME SEVERITY, at vi (1985).

26. For example, "A person plants a bomb in a public building. The bomb explodes and 20 people are killed." *Id.*

27. *Id.*

28. *See id.* at vi–x; *see also* Alfred Blumstein & Jacqueline Cohen, *Sentencing of Convicted Offenders: An Analysis of the Public's View*, 14 LAW & SOC'Y REV. 223, 223 (1980) (concluding that there was "considerable agreement across various demographic groups on the *relative* severity of the sentences to be imposed for different offenses, but disagreement over the *absolute*

Such vast ranges in estimates of the seriousness of different offenses are not consistent with the claims that intuitions regarding crime seriousness are simplistic or only generalized.

D. COUNTERARGUMENTS

Although it seems unlikely, it is theoretically possible that some of the results discussed above are due not to systematic nuanced judgments but rather to "noise" or "guessing" on the part of subjects. For example, people assigning offenses to categories could assign them randomly, rather than systematically, making judgments only appear nuanced. There is substantial evidence against this possibility. If individuals were responding randomly, their answers would be inconsistent when they were asked the same question in different ways or on different occasions.[29]

A good example of this consistency is Walker's paired comparison task. By having subjects judge which in each pair of offenses is more serious, they are, in essence, answering the same question multiple times. If people are responding randomly, there should be many cases in which A is rated more seriously than B, and B more than C, but C more than A. Such "intransitive" groupings would be evidence of random responding. However, paired comparisons yielded few intransitive groupings—less than five percent aggregated across the groups of subjects.[30] This result suggests that variability in judgments is systematic, rather than random.[31]

magnitude of these sentences").

29. As an example, consider subjects asked to judge how "glorky"—a made-up term—a set of objects are. Subjects might assign arbitrary values, perhaps even using a wide range. However, if they came back two months later, they would be unable to assign the same values except by remembering the values they originally assigned. This procedure, referred to as test-retest, can be measured using an index that ranges between -1 and +1 (with 0 meaning no relationship, +1 meaning perfect correspondence from time 1 to time 2, and -1 meaning an inverse relationship). A conceptually similar method is to have subjects rate the same items using different methods and compare the results.

30. *See* Walker, *supra* note 17, at 350–55 (comparing relative rankings and data across social status and gender, while ignoring very severe, and very minor sentences, category and ratio scales).

31. Another way to estimate inability to make discriminations is to look at cases in which people report "don't know" for paired comparisons. This is a direct estimate of the number of cases in which two items are below the subject's ability to make a discrimination. In this case, the average number of "don't knows" was three percent or less across groups. *Id.*

Similarly, some studies have subjects do the task using more than one method—for instance, magnitude estimation and assigning offenses to categories. If subjects are assigning answers randomly, the relationship between the results on these tasks should be small. In contrast, when subjects used multiple methods to rate offenses in the National Crime Survey,[32] there was nearly perfect correspondence between tasks.[33]

A second kind of evidence that people are not just guessing stems from the data regarding consensus. If people are simply guessing or assigning numbers or categories to offenses randomly, there should be little consensus. In fact, as the evidence reviewed in the next sections documents, there is much consensus. In addition, measures of variation in the assignment of offenses to categories or to magnitudes should be high if people were simply assigning these randomly. These measures from the studies reviewed are much smaller than random responding would produce.[34]

32. *See supra* notes 25–26 and accompanying text.

33. *See* V. Lee Hamilton & Steve Rytina, *Social Consensus on Norms of Justice: Should the Punishment Fit the Crime?*, 85 AM. J. SOC. 1117, 1128–42 (1980) (comparing test results with the results from another task in which seriousness was evaluated by assigning a given offense to a set of lines, with longer lines corresponding to greater seriousness).

34. For a relevant measure of variability, see Walker, *supra* note 17, at 351–53. Rossi et al., *supra* note 19, at 227–31, found reasonably low measures of variance. The values for "variance" appear relatively large in some cases, but this is because variance, instead of standard errors, was reported. *Id.* Because each value derives from at least one hundred subjects, the standard error must be one-tenth or less of the reported variance. So even when the variance is large (e.g., 7.479 for "Using LSD"), the standard error of the mean value must be less than 0.748. *Id.* at 227.

Sellin and Wolfgang report standard errors of the mean of (logarithmically) transformed values in their study of offense seriousness. *See* SELLIN & WOLFGANG, *supra* note 16, at 259–73. Whether these error values are "large" or "small" is subjective, but they are generally between 0.01 and 0.02, which seems reasonably to be characterized as "small." *Id.*

Jeffrey Roth's study with prosecutors yielded similar findings. Looking only at the limited list of thirteen offenses, the magnitude estimate scale showed over an order of magnitude of difference among offenses and striking similarities across jurisdictions. Jeffrey A. Roth, *Prosecutor Perceptions of Crime Seriousness*, 69 J. CRIM. L. & CRIMINOLOGY 232, 235 (1978). Roth reported standard errors rather than variance and these were reasonably small, with some exceptions (e.g., "assault—no injury" showed extreme variation in some jurisdictions). *Id.* at 235, 242 ("These results are consistent with previous evidence that the crime seriousness scale is invariant with respect to a wide variety of geographical and personal characteristics.").

E. SMALL FACTUAL CHANGES PRODUCING SIGNIFICANT DIFFERENCES IN LIABILITY JUDGMENTS

Another indication of the nuanced nature of intuitions of justice is the fact that small factual differences among scenarios produce significant differences in the amount of punishment that subjects think is deserved.[35] The studies reported in Paul Robinson and John Darley's study, *Justice, Liability and Blame: Community Views and the Criminal Law*, illustrate the point.[36] In each of eighteen studies, participants were given a series of scenarios that followed the same basic story but varied the facts in a variety of ways.[37] Participants were then asked to give deserved liability and punishment judgments for each scenario.[38] Changes in responses among the scenarios could then be attributed to the varied facts.[39] The results dramatically and repeatedly illustrate how small changes in facts can produce large differences in the punishment judged to be deserved on an astounding array of criminal law issues.[40]

In a study concerning the objective requirements for attempt liability, participants dramatically alter the liability and punishment they would impose according to just how far the actor goes toward committing a coin store robbery.[41] When the person visits the store to plan the robbery, has a special tool made that he will need, and tells friends of his plan, but is caught before he goes to the store to rob it, his average punishment is just under two weeks,[42] but if he takes the additional step of going to the store and starting work on the safe, his liability jumps to 1.7 years.[43] If he voluntarily renounces his attempt before being caught, 85% or more of participants would impose no punishment, but if he voluntarily renounces after committing the offense, few would give him a defense—4%

35. These quite nuanced and sophisticated intuitions exist across demographics, and, as will be apparent in the next Part, there is a high degree of agreement on a wide variety of issues. *See infra* Part II.

36. *See* ROBINSON & DARLEY, JUSTICE, LIABILITY, AND BLAME, *supra* note 1, at 50–51, 79–81, 123–25, 197–99.

37. *Id.* at 7.

38. *Id.*

39. *See id.* at 7–11.

40. *Id.* at 50–51, 79–81, 123–25, 155, 197–99.

41. *Id.* at 20 tbl.2.2.

42. *Id.* (scenario 3).

43. *Id.* (scenario 5).

would allow a defense, and only 27% would allow a defense even if he completely undoes the harm of the offense.[44]

In a study concerning objective requirements for complicity liability, participants dramatically alter the liability and punishment they would impose according to how much assistance the accomplice provides.[45] The man who helps a woman in planning the killing of her husband by directing her to a gun store is given an average liability of five years, while the man who helps her by giving her his gun so she does not have to go to the store gets an average of life imprisonment.[46] But if he offers her his gun but she says she does not need it because she has already killed her husband, then 85% would impose no punishment.[47]

Another study tests the effect of various conditions on punishment for an omission to act that results in a death.[48] The person who fails to throw a stranger a life preserver gets punishment of seven and a half weeks, but if the person is a security guard for the pier, his liability jumps to 3.8 years.[49] Moreover, if the person fails to jump in and save the stranger, knowing that sharks inhabit that body of water, 86% would impose no punishment.[50]

In a study that examines the use of deadly force in self-defense, most people impose no punishment if such force was unavoidable for defense, but impose an average punishment of 9.6 months if the person could have safely retreated, unless the person is in his own house, in which case there is no punishment even if the person could have safely retreated.[51] If he is in a public place but mistakenly believes that he cannot retreat, then he gets no punishment.[52] But if his mistake is not about whether he can safely retreat but about the legality of using deadly force in the situation, then his average punishment is 5.8 years.[53] Interestingly, most of the distinctions that partici-

44. *Id.* (compare scenarios 3a and 5a with scenarios 6a and 6b).
45. *Id.* at 36 tbl.2.9.
46. *Id.* (compare scenarios 4 and 5).
47. *Id.* (scenario 2).
48. *Id.* at 42–43.
49. *Id.* at 45 tbl.2.12 (scenarios 1 and 3).
50. *Id.* (scenario 5).
51. *Id.* at 56 tbl.3.1 (scenarios 2, 4, and 5).
52. *Id.* (compare scenarios 4 and 7).
53. *Id.* (scenario 9).

pants make track criminal law doctrine, even though participants generally do not know the criminal law rules.[54]

Another study examined the effect of mistake in various contexts.[55] Where a person burns a house he thinks he just bought, his punishment will be 1.6 years if he disregards a risk that the house has not yet been legally conveyed to him (which turns out to be the case), but his punishment is only 4.4 months if the risk never occurs to him but would have occurred to the reasonable person.[56] But if he disregards the risk that ownership has not yet been conveyed to unimproved property, rather than to a house, then his punishment is nine months rather than 1.6 years.[57] And if the risk that the unimproved property is not legally his never occurred to him but would have occurred to the reasonable person, then his liability is 2.4 months rather than nine months.[58]

Similarly, if a person has consensual intercourse with a partner who he knows is under the legal age, then his average punishment is 1.6 years, but if he believes she is over the legal age, but is aware of a possibility that she might not be, then his punishment is six months.[59] If her being underage never occurs to him but would have occurred to the reasonable person, then his punishment is six days.[60] If the reasonable person would not have thought she was underage, then he has no punishment.[61]

In a study concerning coercion and entrapment, a truck driver gets two years for voluntarily delivering drugs, but only 4.8 months where he is coerced to make the delivery by his employer's threat to fire him if he does not.[62] But where he makes the delivery because his family was threatened, his punishment is only about four weeks.[63]

54. *Id.* at 201–15 (discussing the extent to which lay intuitions track or deviate from criminal law rules).

55. *Id.* at 84–96.

56. *Id.* (compare scenarios 3b and 3c).

57. *Id.* (comparing the results of scenarios 4b and 3b).

58. *Id.* (scenario 4c).

59. *Id.* (compare scenarios 6a and 6b).

60. *Id.* (scenario 6c).

61. *Id.* (scenario 6d).

62. *Id.* at 151 tbl.5.7 (scenarios 1 and 3).

63. *Id.* (scenario 2).

One study concerns accomplice liability for one perpetrator killing another during a store robbery.[64] Where the accomplice agrees beforehand to the shooting, his punishment is thirty years to life when the principal shoots the store owner, but if the accomplice thought the principal's gun would be unloaded, his liability is only 6.6 years for the owner's death.[65] And if the principal shoots a co-felon rather than the store owner, then the surviving co-felon's punishment is only twelve months.[66]

Another study examines the effect on liability and punishment of different causal chains between a stabbing and a resulting death.[67] Where the death is immediate, the average penalty is life imprisonment.[68] But if death results because a careless nurse at the hospital injects the wrong medication during treatment, the attacker's punishment is only 19.5 years, and if death results because the victim flees the attack and happens to be running under a construction crane when its cable breaks, then the attacker's punishment is 10.2 years[69]—about the same as it would be for an attempted murder under similar circumstances.[70] If the victim's death results from a rare allergic reaction to a common drug during treatment at the hospital, then the attacker's punishment is about twenty-six years, rather than life,[71] but if death results later from an auto accident on the way to the hospital for post-operative treatment, then the attacker's punishment is only fifteen years.[72]

Joseph Jacoby and Francis Cullen's study also found substantial variation in preferred punishment depending on a number of factors—e.g., harm caused, dollar loss, etc.—again supporting the view that punishment preferences are nuanced.[73] However, it is important to note that a factor that mediates seriousness for one crime does not necessarily do so for

64. *Id.* at 169–81.
65. *Id.* at 172–73 tbl.6.3 (compare scenarios 4 and 6).
66. *Id.* (scenario 9).
67. *Id.* at 181–89.
68. *Id.* at 183 tbl.6.7 (scenario 1).
69. *Id.* (compare scenarios 5 and 7).
70. *Id.* (drawing on the results of scenario 2).
71. *Id.* (comparing the results of scenarios 1 and 4).
72. *Id.* (scenario 6).
73. Joseph E. Jacoby & Francis T. Cullen, *The Structure of Punishment Norms: Applying the Rossi-Berk Model*, 89 J. CRIM. L. & CRIMINOLOGY 245, 271–86 (1998).

another.[74] For example, the dollar amount stolen changes seriousness judgments for theft, but not for theft in the context of rape.[75] That the effect of mediating factors depends on the type of crime again indicates extremely nuanced intuitions.

Changes in the characteristics of the offender also influence judgments. In a study concerning immaturity and other excuses, an eighteen-year-old male intentionally sets another boy on fire to kill him while he sleeps, getting an average punishment of 25.5 years.[76] A fourteen-year-old male who does the same thing gets 6.2 years, and a ten-year-old gets eleven months.[77] The adult gets on average a split between life imprisonment and the death penalty.[78] Peter Rossi similarly found that preferred punishments varied depending not only on the consequences of the crime, but also on features of the victims and offenders.[79]

In short, the evidence from multiple testing methods supports the view that intuitions of justice are finely nuanced. Durham summarizes the surveys this way: "Virtually without exception, citizens seem able to assign highly specific sentences for highly specific events."[80] The conclusion suggested by the empirical evidence is that people take account of a wide variety of factors and often give them quite different effect in different situations. That is, people's intuitions of justice are not vague or simplistic, as claimed, but rather sophisticated and complex.

II. IS THERE AGREEMENT AMONG PEOPLE'S INTUITIONS OF JUSTICE?

Even if people have a quite nuanced notion of desert, is there so much disagreement about these notions that they could not be operationalized in a criminal code or in sentencing

74. Gottfredson et al., *supra* note 18, at 29–37.

75. *Id.* at 39 ("The studies described above suggest that we consider not only the type of offense and the amount of loss incurred (measured in dollars) when judging the seriousness of criminal or delinquent acts, but also the interaction between the two.").

76. ROBINSON & DARLEY, JUSTICE, LIABILITY, AND BLAME, *supra* note 1, at 139–47.

77. *Id.* at 141 tbl.5.5 (scenarios 7, 8, and 9).

78. *Id.* (scenario 1).

79. *See* Peter H. Rossi et al., *Beyond Crime Seriousness: Fitting the Punishment to the Crime*, 1 J. QUANTITATIVE CRIMINOLOGY 59, 60–61 (1985).

80. Alexis M. Durham III, *Public Opinion Regarding Sentences for Crime: Does it Exist?*, 21 J. CRIM. JUST. 1, 2 (1993).

guidelines? Many writers have argued that people simply disagree in their notions of justice.

For instance, Michael Tonry argues that "even assuming retribution in distribution is appropriate, there is a classic epistemological problem. How do we know how much censure, or 'deserved punishment,' a particular wrongdoer absolutely deserves? God may know, but as countless sentencing exercises have shown, people's intuitions about individual cases vary widely."[81]

Similarly, John Monahan concludes:

> There is . . . reason to doubt that anything like a consensus exists on the seriousness of criminal conduct. While there may be some agreement on relative levels of harm, there appears to be great variation in perceptions of the absolute magnitude of harm represented by various criminal acts, and in either the relative or absolute level of culpability represented by various criminal actors.[82]

---

81. Michael Tonry, *Obsolescence and Immanence in Penal Theory and Policy*, 105 COLUM. L. REV. 1233, 1263 (2005).

82. John Monahan, *The Case for Prediction in the Modified Desert Model of Criminal Sentencing*, 5 INT'L J.L. & PSYCHIATRY 103, 105 (1982). Other authors have made similar arguments:

> [Desert theorist John Kleinig assumes that] we can work out one single, linear ordering of crimes, from least to most "serious." Yet that scarcely seems a credible assumption. Try, for instance, to rank the following crimes in order of their "seriousness": attempted residential burglary, trading stock on inside information, negligent vehicular homicide, bribing a mine-safety inspector, possessing an ounce of cocaine, and burning a cross on the lawn of black newcomers to a previously all-white neighborhood. To view this motley assortment along a single dimension of "seriousness" would seem no less difficult than to perceive the inner logic behind the apocryphal Chinese encyclopedist of Jorge Luis Borges's imagination.

David Dolinko, *supra* note 12, at 1638–39 (1992). Similarly, van den Haag argues:

> [Desert theorist Andrew von Hirsch] appears to believe that the comparative seriousness of crimes can be determined in all cases. Not so. Comparative seriousness can be determined only for some crimes, and it does not fully determine the comparative punishment deserved. If rape is a crime and murder is a crime, rape-murder must be more serious than either. Does rape-murder deserve the sum of the punishments meted out for rape and for murder? More? Less? Even when crimes are nearly homogeneous, assigning seriousness is arbitrary: Is rape more serious than assault with a deadly weapon? Is burglary more serious than fraud when fraud does more harm? What about mishandling toxic waste? Ordinal determinations of seriousness become altogether arbitrary when the seriousness of heterogeneous crimes must be compared.

Ernest van den Haag, *Punishment: Desert and Crime Control*, 85 MICH. L. REV. 1250, 1254 (1987).

This view, together with the view that people can make only crude rankings of offenses by seriousness, has been used to justify important policy decisions, including the drafting of the U.S. Sentencing Commission Sentencing Guidelines:

> [S]ome students of the criminal justice system strenuously urged the Commission to follow what they call a "just deserts" approach to punishment. . . . The difficulty that arises in applying this approach is that different Commissioners have different views about the correct rank order of the seriousness of different crimes. . . . Considering the inherent subjectivity of such a trade-off process, the Commission soon realized that only a crude ranking of behavior in terms of just deserts . . . could be developed.[83]

But the empirical evidence again presents a different picture. The previous Part has already discussed the quite nuanced nature of intuitions of justice. The empirical evidence also suggests a quite strong agreement on a wide variety of liability and punishment issues across demographics. In Part II. A. this Article examines the most fundamental claim: that humans share an intuition that serious wrongdoing should be punished. Part II. B. considers the extent of the agreement on the relevant seriousness of different instances of wrongdoing.

A. THE INTUITION THAT SERIOUS WRONGDOING SHOULD BE PUNISHED

Using a range of techniques, previous empirical studies confirm a nearly universal human intuition that serious wrongdoing deserves punishment.

1. Questionnaire Studies

In much of the previous research using questionnaires, subjects have had an option to assign no punishment. Yet, people overwhelmingly chose to assign punishment even though they could have assigned none. For example, a study by Craig Boydell and Carl Grindstaff gave participants the opportunity to indicate the penalty they believed should be applied to an offense, as well as the minimum and maximum penalty applicable to that offense.[84] For the least serious offense they investigated, a mere four percent of respondents indicated that the

83. Breyer, *supra* note 12, at 15–17.

84. Craig L. Boydell & Carl F. Grindstaff, *Public Opinion Toward Legal Sanctions for Crimes of Violence*, 65 J. CRIM. L. & CRIMINOLOGY 113, 113, 116 (1974) ("With regard to crime . . . and particularly crimes against the person, it may be that there exists a common denominator such as fear . . . that obscures the socio-demographic interests that normally are important.").

minimum penalty they would apply is "no punishment."[85] For all other crimes, "no punishment" was chosen as the appropriate penalty, the maximum penalty, and the minimum penalty by less than four percent of the respondents.[86] Indeed, in the majority of cases, "no punishment" was selected by no participants, even as the minimum punishment for the offense.[87]

Similarly, a 1985 study allowed people to indicate zero in their magnitude estimation task when questioned about the appropriate punishment for certain offenses.[88] The frequency of zeros is not reported but the average value assigned for even the offense judged least serious—"a person under 16 years old plays hooky from school"—was statistically different from zero.[89] On average, then, across all regions investigated, even the least serious offense was judged to deserve some punishment.[90]

In studies in which "no liability" was not an option—the majority of studies reviewed here—it is clear that the experimenter assumed that all subjects would believe that all acts described in the study deserved some punishment.[91] This assumption itself is noteworthy. Researchers from multiple disciplines over the course of decades have routinely assumed that no significant fraction of subjects would believe that the offenses in question deserved no punishment. In short, even the experts in these fields discount the possibility that subjects would not want to punish. If all of these researchers are wrong, there likely would be evidence in the studies of subjects refusing to assign punishment, by responding with only the minimum possible amount of punishment in each case or by responding randomly. Yet these types of results did not occur, vindicating the researchers' views.

85. *Id.* at 114 tbl.1. Note that the penultimate column is mislabeled "education" and should read "execution"; therefore selections in this column should not be interpreted as a preference for "no punishment." *See id.* at 115.

86. *See id.*

87. *See id.*

88. WOLFGANG ET AL., *supra* note 25, app. A at 137 (instructing respondents: "If YOU think something should not be a crime, give it a zero").

89. *See id.* at 46–47 tbl.1.

90. *See id.* app. C at 158–61 tbl.C-1; *see also* Gibbons, *supra* note 19, at 395. For each of twenty offenses, there were roughly three hundred respondents. *Id.* Only four of the twenty offenses yielded more than 10% "no penalty" responses. *See id.* tbl.1. These were homosexuality, statutory rape (consensual sex between a twenty-year-old and a sixteen-year-old), draft evasion, and marijuana use. *Id.* at 394, 395 tbl.1.

91. WOLFGANG ET AL., *supra* note 25, at 47 tbl.1.

Taken together, these facts suggest that every study conducted, including the ones reported here, supports the view that people share the intuition that serious wrongdoing should be punished.

2. Behavioral Economics Studies

In questionnaire studies, people are asked how much punishment is "right" or "justified" or "deserved." They are not, however, required to incur any cost to inflict punishment. In some experiments, however, people must actually bear a cost to punish.[92] These experiments typically investigate breaches of perceived norms, such as fairness, rather than criminal offenses.[93] That is, the wrongdoing being punished here typically is dramatically less serious than the kind of wrongdoing involved in criminal violations.[94] Thus, to the extent that persons are demonstrated as willing to bear the cost of punishment for violation of a civil norm, it seems reasonable to assume that they would similarly be willing to bear the cost of punishment—or pay much more to punish—in cases of criminal wrongdoing.

One game used in such studies tests the willingness to bear the costs to punish perceived unfairness is the so-called Ultimatum Game.[95] In the typical version, experimental subjects are brought into the laboratory and randomly assigned to one of two experimental "roles," either that of the Proposer or the Responder.[96] The Proposer is provisionally allocated a sum of money, called an "Endowment," often ten dollars.[97] The Proposer suggests a split of the Endowment with the Responder, for example, six dollars for the Proposer, four dollars for the Responder.[98] The Responder is then given the option of accepting the offer, in which case the money is split as designated by

92. *See* Daniel Kahneman et al., *Fairness and the Assumptions of Economics*, 59 J. BUS. S285, S290–91 (1986) (documenting how study participants voluntarily incurred a penalty in order to punish others for unfair behavior).

93. *See id.* at S288–92 (describing an experiment where the violation of the fairness norm was the inequitable distribution of twenty dollars between two participants).

94. *Id.*

95. *See* COLIN CAMERER, BEHAVIORAL GAME THEORY: EXPERIMENTS IN STRATEGIC INTERACTION 8–12, 43–113 (2003).

96. *Id.* at 8.

97. *Id.*

98. *Id.*

the Proposer, or rejecting the proposal, in which case the ten dollars is not given to the subjects.[99]

Proposers generally offer between 40% and 50% of the Endowment to Responders,[100] but our interest is in situations in which Proposers offer a very unequal split. Under these conditions, Responders often reject the proposals, costing them the amount offered by the Proposer, and thus depriving the Proposer of her portion of the money.[101] Such rejections are interpreted by researchers as cases in which Responders are punishing Proposers for making unfair offers.[102] This punishment happens under carefully controlled conditions, when the subjects do not physically interact with one another, do not know one another's identities, and when even the experimenter does not know the Responder's decision.[103] In short, people punish perceived unfairness at a cost to themselves, even when there are no instrumental consequences or experimenter expectations that might be at work.[104] Even more striking, there is evidence that people will pay to punish another person even when they are not directly involved in the transaction if they perceive that the person has behaved intentionally unfairly.[105]

99. *Id.* It is important to note that these experimental games are almost always played for real money.

100. This result varies considerably depending on the details of the experimental procedure. *See id.* at 50–52 tbl.2.2.

101. *Id.* at 53–55 tbl.2.3.

102. *Id.* at 10.

103. *See* Gary E. Bolton & Rami Zwick, *Anonymity Versus Punishment in Ultimatum Bargaining*, 10 GAMES & ECON. BEHAV. 95, 111 (1995) (showing that punishment occurs even when experimenters do not know subjects' decisions).

104. A wealth of data from behavioral economics, using methods from other games as well, is consistent with this conclusion. CAMERER, *supra* note 95, at 43–117, is an excellent source for relevant work. For recent cross-cultural work using the Ultimatum Game, see Joseph Henrich et al., *"Economic Man" in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies*, 28 BEHAV. & BRAIN SCI. 795, 799–801 (2005).

105. Kahneman et al., *supra* note 92, at S288–92. One set of subjects played a Dictator Game, which is similar to the Ultimatum Game except that the Responder must accept any offer. *Id.* at S290. After the Dictator Game was played, a separate set of subjects was given the opportunity to pay one dollar to deprive five dollars from someone who had taken a very unequal split (eighteen dollars / two dollars) in an earlier Dictator Game. *Id.* at S290–91. A "clear majority (74%)" of subjects chose to do so. *Id.* Although subsequent work has suggested that this experiment might have overestimated "third party punishment," it is nonetheless clear that in the correct circumstances, people will pay to punish those who are perceived to have violated a fairness norm. *See, e.g.*, Ernst Fehr & Urs Fischbacher, *Third-Party Punishment and Social*

3. Cross-Cultural Studies

If notions of wrongdoing and punishment were absent from a culture, study respondents presumably would have difficulty understanding the task put to them or would be expected to randomly answer questions about these topics. We know of no researcher who has attempted to gather data on judgments about wrongdoing or punishment who has had difficulty in conveying the relevant questions and in obtaining coherent responses.[106]

In fact, cross-cultural data suggest that questionnaire studies yield similar results in all of the cultures that have been studied. While we do not deny that there are important cultural differences, the intuition that those who commit wrongs should be punished seems to be universal. We do not know of any documented case of any culture in which the intuition to punish serious wrongdoing did not exist.

Experts commenting on this element of human culture show broad agreement. Cultural psychologist Paul Rozin and his colleagues conclude that "[m]oral judgment and the condemnation of others, including fictional others and others who have not harmed the self, is a universal and essential feature of human social life."[107] Similar sentiments have been expressed by developmental psychologist Jerome Kagan, who includes this intuition as one of "a limited number of universal moral categories that transcend time and locality."[108] Philosopher Ray Jackendoff concludes: "Thus in our culture, the legal system punishes not only physical aggression like assault, but also economic aggression like stealing. Similar institutions are found in some form in every culture, even in the absence of written legal codes."[109] Anthropologist Donald Brown, in his exhaustive re-

Norms, 25 EVOLUTION & HUM. BEHAV. 63, 68 (2004) ("Most third parties punished dictators who transferred less than half their endowment, and the majority of recipients expected them to do so.").

106. *See* John R. Snarey, *Cross-Cultural Universality of Social-Moral Development: A Critical Review of Kohlbergian Research*, 97 PSYCHOL. BULL. 202, 226 (1985) (reviewing cross-cultural data that reveal important universals in moral development).

107. Paul Rozin et al., *The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity)*, 76 J. PERSONALITY & SOC. PSYCHOL. 574, 574 (1999). For a discussion of some psychologists (and others) who want to abolish punishment, see STEVEN PINKER, THE BLANK SLATE 181 (2002).

108. JEROME KAGAN, THE NATURE OF THE CHILD 118–19 (1984).

109. RAY JACKENDOFF, LANGUAGE, CULTURE, CONSCIOUSNESS: ESSAYS ON

view of the cross-cultural data, included intuitions surrounding justice and punishing transgressors as a "Human Universal."[110] In short, experts from multiple disciplines have unambiguously asserted that, despite cultural differences, the intuition to punish is a key aspect of what it means to be a member of the human species.

4. Are There Exceptions?

We do not mean to imply that no human could conceive of a social structure in which people were not punished for their transgressions. We note that some religious writings might be taken to urge such a view, such as the biblical injunction to "turn the other cheek." But we suggest that the significant point here is that this injunction is given precisely because it is understood to run counter to strong intuitions to punish serious wrongdoing.[111] More importantly, it is hardly evidence contradicting the existence of shared intuitions of justice that some people have called for the abolition of punishment, if the call is rarely if ever answered.

Of course, one might argue that there are in fact historical figures who have believed that people should not be punished for their wrongdoing. For example, some might argue that Jesus Christ or Mahatma Gandhi were persons who would in fact "turn the other cheek," disproving the claim of a universal human intuition to punish serious wrongdoing. There is reason to believe that even such icons of love and forgiveness in fact believed in punishment of wrongdoing. Biblical passages of

MENTAL STRUCTURE (forthcoming June 2007) (manuscript at 4-25, available at http://people.brandeis.edu/~jackendo/ch4Soccog.pdf).

110. DONALD E. BROWN, HUMAN UNIVERSALS 138 (1991). In describing the Universal People (UP)—his term for those features that all people from all cultures share in common—he writes:

> The UP have law, at least in the sense of rules of membership in perpetual social units and in the sense of rights and obligations attached to persons or other statuses. Among the UP's laws are those that in certain situations proscribe violence and rape. Their laws also proscribe murder—unjustified taking of human life (though they may justify taking lives in some contexts). They have sanctions for infractions, and these sanctions include removal of offenders from the social unit—whether by expulsion, incarceration, ostracism, or execution. They punish (or otherwise censure or condemn) certain acts that threaten the group or are alleged to do so.

Id.

111. One translation of the relevant passage is: "I say unto you, That ye resist not evil: but whosoever shall smite thee on thy right cheek, turn to him the other also." *Matthew* 5:38–:39 (King James).

Christ's teachings are full of references calling for punishment.[112] And one might observe that Mahatma Gandhi advocated strikes and sit-ins meant to punish the British.[113] But the important point here is not the conduct of these persons and others like them but rather the fact that they have become historic figures precisely because they seemed to advocate a position so difficult for ordinary persons and so counter to human nature. Their historic status only serves to illustrate and confirm the strength of the human intuition to punish serious wrongdoing.

B. INTUITIONS ON THE RELATIVE SERIOUSNESS OF WRONGDOING

As is apparent from the quotations at the start of this Part, some writers might concede the existence of a shared intuition that serious wrongdoing should be punished but might dispute nonetheless that there is agreement on how punishment ought to be distributed across cases of wrongdoing. However, a substantial body of research indicates a broad consensus regarding the relative seriousness of different wrongdoings and the appropriate relative amount of punishment.[114]

It is not that everyone agrees on a specific sentence for each case. On the contrary, some people would give generally harsher punishment and others generally less harsh punishment. But whether harsh or lenient punishers, people tend to

112. *See, e.g.*, *Matthew* 25:46 (King James) ("These shall go away into everlasting punishment."); 2 *Thessalonians* 1:5–:10 (King James) ("Who shall be punished with everlasting destruction from the presence of the Lord, and from the glory of his power. . . ."). Note that Paul does not seem to shy away from punishment in the present. He indicates that an individual in power "beareth not the sword in vain: for he is the minister of God, a revenger to execute wrath upon him that doeth evil." *Romans* 13:4 (King James).

113. Gandhi used civil disobedience as a tool to induce the British to leave. Civil disobedience was certainly a cost imposed on the British, particularly as they were fighting World War II. MAHATMA GANDHI, THE ESSENTIAL GANDHI: AN ANTHOLOGY OF HIS WRITINGS ON HIS LIFE, WORK, AND IDEAS 121–27 (Louis Fischer ed., 2d ed. 2002). Punishment is inflicting a cost, albeit in this case on a group, rather than an individual.

114. An important distinction in discussion of consensus is between absolute and relative judgments of seriousness or desert. Consensus in absolute judgments would mean that people agree on the seriousness of punishment—fine, sentence length, none, etc.—for a given offense. Consensus in relative seriousness means that people agree on the ordering of offenses such that they agree which offenses are more serious than others but not necessarily on the specific amount of punishment deserved. *See* SELLIN & WOLFGANG, *supra* note 16, at 268.

agree on the relative degree of blameworthiness among a set of cases. That is, while they may disagree as to the point to which the punishment continuum should extend at its high end, they agree on the relative placement of cases along that continuum. Once a society determines the end point of the punishment continuum, shared intuitions of justice will set each case on a specific point on the continuum in its appropriate place relative to other cases. The specific amount of punishment due each case is fixed, then, not because there is some magical connection between that amount of punishment and that particular offense but rather because that is the amount of punishment needed to distinguish that case from cases of noticeably greater and lesser blameworthiness on the limited continuum of punishment.[115]

The existing studies reveal an extraordinary extent of agreement across a variety of issues and demographics. While a variety of testing techniques and subjects have been used, the conclusions are all essentially the same, confirming the existence of shared intuitions as to relative seriousness of different variations on wrongdoing.

1. Previous Domestic Studies

The most well known study is that of Sellin and Wolfgang, who in the 1960s surveyed 575 individuals across Pennsylvania.[116] Subjects were asked about the seriousness of fifty-one offenses, such as: "An offender forces a female to submit to sexual intercourse. No other physical injury is inflicted."[117] The researchers used two test methods. One method asked subjects to

115. *See supra* text accompanying notes 10–11. When the U.S. Sentencing Commission reverted to basing its guidelines on mathematical averaging of past sentencing practice, it ignored the importance of this difference between absolute and relative punishment amounts. *See* Breyer, *supra* note 12, at 17. If it had sought to rely upon what judges had done in the past, it ought to have adjusted the punishment a judge handed out in different cases to account for whether that judge was generally more or less severe than the average judge. Different judges tended to contribute different kinds of cases to the statistical pool, thus each judge would likely have the greatest influence on the guidelines for the kind of cases she contributed.

116. *See supra* notes 22–23 and accompanying text; *see also* SELLIN & WOLFGANG, *supra* note 16, at 338–39 (describing a brief history of the limited work measuring perceptions of crime seriousness up to this study).

117. SELLIN & WOLFGANG, *supra* note 16, at 381, 391. Twenty-one base offenses were given to each rater in random order, plus a random assortment of thirty other crimes drawn from the remaining 120 offenses. *Id.* at 277.

place offenses on a scale ranging from one to eleven.[118] Another method, referred to as a "magnitude estimation task," asked subjects to assign the offense a number to indicate its seriousness relative to bicycle theft, which was assigned a value of ten.[119] The results show broad agreement.[120] The researchers conclude that "[t]he most strongly supported conclusion . . . is that all the raters . . . tended to assign the magnitude estimations [so] that the seriousness of the crimes is evaluated in a similar way, without significant differences, by all the groups" and, further, that a "pervasive social agreement about what is serious and what is not appears to emerge . . . ."[121] This study has been replicated, with results that support its original conclusions.[122]

Since Sellin and Wolfgang, there have been many other studies using a variety of methods, all reaching similar conclusions. Here is a sampling of five that are representative.[123] Jacoby and Cullen surveyed a national sample of 1920 adults who were read eight crime vignettes during thirty-minute telephone interviews.[124] The type of crime in each vignette was

118. *Id.* at 254.

119. *Id.*

120. For scale category means broken out by groups, see *id.* at 387–88.

121. *Id.* at 268.

122. For example, Roth replicated the Sellin and Wolfgang study in 1978, surveying 909 prosecutors across the United States. *See* Roth, *supra* note 34, at 233–34. Subjects were given a list of thirty-six brief offense descriptions, just as in the Sellin and Wolfgang survey. Thirteen offenses were the same for every survey; the rest were randomly selected from the remaining 250 offenses. A ten-dollar larceny was the first crime in every booklet; the rest of the crimes were randomized. All items were drawn from the Sellin-Wolfgang index. Those surveyed were asked to estimate the relative seriousness of crimes. *Id.* at 233–35. Roth concluded, "These results are consistent with previous evidence that the crime seriousness scale is invariant with respect to a wide variety of geographical and personal characteristics." *Id.* at 242.

123. In a survey by Boydell and Grindstaff, 451 household heads in London, Ontario, were given a list of eight crimes against a person—murder, rape, robbery, etc.—and asked the penalty they would impose for a typical offense, as well as the minimum and maximum penalty that ever should be imposed for the offense. Boydell & Grindstaff, *supra* note 84, at 113–14. The researchers analyzed the penalties imposed for each of the eight offenses according to the sex, age, income, and religious differences among the subjects and found few statistically significant differences. *Id.* at 116. They reported, "Of the forty-eight relationships investigated, only three (all relating to religion and religious attendance) produced statistically significant differences, and no general trends were discernible." *Id.* They concluded, broadly, that demographics do not have important effects on people's views on punishment for crimes against the person. *Id.*

124. *See* Jacoby & Cullen, *supra* note 73, at 263–65. One scenario included

taken from a list of twenty-four offenses.[125] "Respondents agreed, on average, on the ordering of the twenty-four crimes in seriousness and deserved sentence length."[126] The researchers reported that the "amounts of variation in sentencing accounted for by the individual characteristics studied was quite small, less than 10%."[127]

Blumstein and Cohen surveyed 603 residents of Allegheny County, Pennsylvania; subjects were asked to assign the length of a prison sentence[128] that "best fits the seriousness of the offense" for twenty-three offenses that researchers presented in the form of brief crime scenarios.[129] The researchers found no strong effects of demographics—including gender, race, religious affiliation, and level of education—on the ordering of sentences. That is, different groups tended to agree on which crimes should be punished more than other crimes. They concluded that there was "considerable agreement across various

a bicycle theft, again assigned a seriousness of ten. *Id.* at 268–69. The seven non-bicycle vignettes were different for each subject and were constructed by a computer that randomly picked alterative facts relating to thirteen different aspects of the offense, including "the type of crime, amount of harm incurred by the victim, offender characteristics, and victim characteristics." *Id.* at 268. Subjects also were asked to select the sentence they would give each offender. *Id.*

> All the commonly available punishments—jail or prison, probation, fine, restitution, and (for homicide offenses only) death—were then offered. Respondents were asked which of these punishment types they would choose for the offender in that crime vignette. If they chose incarceration, they were asked whether the time should be served continuously or periodically and how long the sentence should be. If they chose a fine, they were asked the amount. Respondents could choose as many of these punishment types as they wished for each vignette.

Id. at 269.

125. *Id.* at 265.

126. *Id.* at 285 ("[The subjects] did not agree on the appropriate value for seriousness or sentence length.").

127. ROSSI & BERK, *supra* note 7, at 205. Translating quantitative results is always problematic, as people can reasonably disagree about what constitutes "a lot" or "a little" disagreement. In this case, the characterization of this level of disagreement—only 10%—as "small" seems very reasonable.

128. Blumstein & Cohen, *supra* note 28, at 227–31. According to the survey instructions, "sentence" meant the amount of time the offender would actually serve. *Id.* at 229.

129. *Id.* at 228–29. A sample scenario: first-degree murder: "The offender is convicted of first degree murder after he intentionally killed a person who witnessed a crime he had committed." The offenses represented a significant contribution to present prison populations in the United States (i.e., first-degree murder, narcotics sales). *Id.* at 228.

demographic groups on the *relative* severity of the sentences to be imposed for different offenses."[130]

Peter H. Rossi, Emily Waite, Christine E. Bose, and Richard E. Berk interviewed 125 whites and seventy-five blacks in Baltimore, Maryland with a roughly equal number of males and females, asking people to categorize eighty offenses each into nine categories according to how serious the offense was perceived to be.[131] The correlations between ratings of blacks and whites, males and females, and more and less educated groups were 0.89, 0.94, and 0.89 respectively, indicating a substantial amount of agreement.[132] The researchers concluded that "the amount of consensus among subgroup averages is impressive"[133] and that "the norms defining how serious various criminal acts are considered to be, are quite widely distributed among blacks and whites, males and females, high and low socioeconomic levels, and among levels of educational attainment."[134]

130. *Id.* at 223 (noting, however, that there was "disagreement over the *absolute* magnitude of these sentences").

These researchers were interested in the relationship between their respondents' assignments of sentences and sentences given to real offenders who had committed crimes that fit the descriptions, as well as the relationship between assignments of sentences and the amount of time served by individuals who had committed such crimes. *Id.* at 258–61. They concluded that "[c]onsiderable discrepancies exist between recommended sentence lengths and actual time served, but there is substantial agreement between recommended sentences and actual sentences." *Id.* at 258.

131. Rossi et al., *supra* note 19, at 226. Sample offenses ranged from the "planned killing of a policeman" to "being drunk in public places." *Id.* at 228–29 tbl.1.

132. *Id.* at 230 tbl.2. Again, characterizing quantitative data as large or small is open to reasonable disagreement. These numbers would generally be considered to be large by conventional standards in statistics. *See* JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 80 (2d ed. 1988) ("[T]he practical upper limit of predictive effectiveness . . . [is] . . . a validity coefficient of the order of .50." (citing E.E. GHISELLI, PERSONAL PSYCHOLOGY 17, 61–63 (1964))). Correlation values range between -1 and +1. A value of -1 means a perfect inverse relationship (as one variable goes up, the other goes down), a value of zero means no relationship (the two variables are unrelated to one another), and a value of +1 means a perfect positive relationship (as one variable increases, the other variable increases). The values reported here, which are close to one, indicate that the ratings of the two groups were very similar to one another.

133. Rossi et al., *supra* note 19, at 231.

134. *Id.* at 237. The analysis also showed that "the more highly educated and the younger respondents were, the more likely were their individual ratings of criminal acts to agree with the average computed for the entire sample . . . suggesting that exposure to the normative structure and language han-

Lee Hamilton and Steve Rytina conducted face-to-face interviews with 391 subjects in the Boston area in which they asked subjects to rank each of seventeen offenses using a "magnitude estimation task" similar to the one described above.[135] A comparison of the individuals' judgments of seriousness and desired punishments with the sample's average judgments gave high correlations—0.71 and 0.73, respectively—suggesting "a high level of consensus."[136] An analysis of the demographic differences among the subjects—age, race, income, and sex—showed no strong effects.[137]

Charles W. Thomas, Robin Cage, and Samuel C. Foster surveyed 3334 households, asking subjects what they felt would be a "fair sentence" for each of seventeen offenses.[138] They reported finding "evidence of a remarkable level of consensus, even after separating the sample on the basis of their sex, race, age, income, occupational prestige, and educational attainment."[139] They concluded that the findings, "regardless of the type or category of offense examined, are not supportive of any prediction that suggests variations between different categories of the population in either perceptions of relative seriousness of these offenses, or the level of sanctions that are viewed as appropriate."[140]

To conclude, we have reviewed studies that asked subjects about crime seriousness, preferred sentences, or both.[141] These

dling ability lead to better knowledge of the normative structure." *Id.*

135. Hamilton & Rytina, *supra* note 33, at 1124–26. Offenses on the list included, for example, "Taking $50" and "Forcible rape." *Id.* at 1125 tbl.1. Subjects were also given a questionnaire that outlined "concrete crimes." *Id.* For example, one such crime was relayed as follows: "One week he [she] was short on money, he [she] took an envelope . . . [that] contained $50 in cash." *Id.* at 1125 n.9. Also included were the appropriate legal charges, for example, "charged with $50 larceny." *Id.* Participants were asked to assess the relative seriousness of these acts and to assign punishments. *Id.*

136. *Id.* at 1132.

137. Hamilton and Rytina report only a "weak" effect of race and income. *Id.* The results of the analysis for education suggest these differences are not due to differences in education by race or income. *Id.* at 1134.

138. Charles W. Thomas et al., *Public Opinion on Criminal Law and Legal Sanctions: An Examination of Two Conceptual Models*, 67 J. CRIM. L. & CRIMINOLOGY 110, 112 (1976). Offenses were described by labels such as "murder" or "rape." *Id.* at 113 tbl.1.

139. *Id.* at 116.

140. *Id.*

141. *See, e.g.*, Blumstein & Cohen, *supra* note 28, at 226 (analyzing Arnold M. Rose & Arthur E. Prell, *Does the Punishment Fit the Crime?*, 61 AM. J. SOC. 247 (1955)); *see also* Hamilton & Rytina, *supra* note 33, at 1130; Michael

two kinds of inquiries are related but not identical. The second includes a broader range of factors—beyond the seriousness of the criminal act itself—such as the consequences of the act, features of the perpetrator, the perpetrator's prior history, and characteristics of the victim.[142] Nonetheless, testing of both offense seriousness and deserved punishment consistently shows substantial consensus.

What is particularly striking about these studies is that, by virtue of their design, one could expect a fair amount of disagreement. When two people are given a skeletal description of an offense, as most of these studies provide, both of the persons are likely to visualize a "story" of the offense, yet they may well "fill in" different details to complete the picture. Demographic differences could prompt subjects to fill in different details, reflecting the different life experiences from which the details are drawn. If subjects visualize slightly different stories, one would expect them to give slightly different assessments of relative seriousness, even if they in fact agreed in their intuitive judgments. This potential for exaggerating the extent of disagreement becomes greater as the crime descriptions become more skeletal, and is at its worst when researchers use crime labels rather than factual descriptions, as in some of the studies reported above, because different people are quite likely to visualize different stories when given just the bare offense label. Despite this serious potential to underestimate the extent of agreement, the studies consistently show a significant level of agreement on intuitions of justice, even across demographics.

Rossi and Berk, in their review of the literature through 1997, suggest that the studies converge on the view that people share intuitions about the relative seriousness of wrongdoing.[143] "[A] [f]airly strong consensus exists on the seriousness

O'Connell & Anthony Whelan, *Taking Wrongs Seriously: Public Perceptions of Crime Seriousness*, 36 BRIT. J. CRIMINOLOGY 299, 310 tbl.2 (1996); Rossi et al., *supra* note 79, at 61.

142. Rossi et al., *supra* note 79, at 63–68 (describing the selection of personal vignettes that include a variety of characteristics unique to the crime, perpetrator, and victim).

143. *See, e.g.*, PETER H. ROSSI & RICHARD A. BERK, U.S. SENTENCING COMM'N, A NATIONAL SAMPLE SURVEY: PUBLIC OPINION ON SENTENCING FEDERAL CRIMES 193 (1995) (discussing sentencing, but in absolute, and not relative terms); Peter H. Rossi & Richard A. Berk, *A Conceptual Framework for Measuring Norms, in* THE SOCIAL FABRIC: DIMENSIONS AND ISSUES 77, 103 (James F. Short, Jr. ed., 1986) (presenting models on consensus and "attempt[ing] to lay the foundations for exploring normative structures based on both technical and conceptual tools"); Peter H. Rossi & Patrick Henry, *Seri-*

ordering of crimes, with those involving actual or threatened physical harm to victims generally considered to be the most serious . . . ."[144] In fact, their summary of previous studies suggest that "there is very little, if any evidence that there exist subgroups within the American population with radically different views about sentencing norms," and that "[t]here is no evidence for a normative order that is an alternative to what the overwhelming majority of the American population believe."[145]

---

ousness: A Measure for All Purposes?, *in* HANDBOOK OF CRIMINAL JUSTICE EVALUATION 489, 489–91 (Malcolm W. Klein & Katherine S. Teilmann eds., 1980); Peter H. Rossi & Richard A. Berk, *Varieties of Normative Consensus*, 50 AM. SOC. REV. 333, 340 (1985) ("Model V more or less describes the majority of, if not most, normative domains in our society: People by and large agree on what the norms are but differ in their degrees of attachment to the normative structure so defined."); *see also id*. at 339 ("Under this condition, individual members of the society need not agree on the specific judgments to be rendered on each moral application, but each individual can be different from the others by [a specific constant].").

144. ROSSI & BERK, *supra* note 7, at 12; *see also* SELLIN & WOLFGANG, *supra* note 16, at 324 (comparing studies of crime seriousness).

145. ROSSI & BERK, *supra* note 7, at 209. Durham expresses skepticism about consensus research. One concern focuses on "the appropriate amount and kind of information to include" when eliciting judgments about specific criminal acts. Durham, *supra* note 80, at 6. This concern is largely irrelevant to our thesis because we are uncommitted with respect to these details, and we accept that information is a potentially important factor in these judgments. Durham also worries about the fact that some stimuli are unfamiliar to subjects. *Id*. (noting that respondents may express an unfounded opinion about an entirely fictitious event). This objection works in our favor; to the extent that subjects are unfamiliar, "noise" is added to the data, and so consensus has been, if anything, underestimated. *See id*. at 1.

In George Bishop and his colleague's test with fictitious stimuli, respondents were divided on an agree/disagree task. George F. Bishop et al., *Pseudo-Opinions on Public Affairs*, 44 PUB. OPINION Q. 198, 199–200 (1980). Lack of familiarity tends to minimize consensus, just as one would expect. *See id*. at 201 tbl.1. The objections that interviews put constraints on time or that respondents might not understand vignettes similarly bolster, rather than undermine our position; such constraints should also add "noise" to the data, undermining consensus. *See* Terance D. Miethe, *Public Consensus on Crime Seriousness: Normative Structure or Methodological Artifact?*, 20 CRIMINOLOGY 515, 523–24 (1982) ("However, it has been argued here that subgroup consensus may not exist at all and if it does, it may be located within specific types of criminal acts due to the possibility of a legal bias in the instructions given to raters. The accurate determination of the extent of public consensus on the seriousness of criminal acts has a number of practical, as well as theoretical, implications. As such, being able to dismiss these alternative explanations is of primary importance for evaluating the utility of seriousness studies. Furthermore, only by adopting the research strategies discussed in this article is it possible to evaluate whether or not consensus truly exists.").

2. Cross-Cultural Studies

The cross-cultural evidence appears to support the view that people everywhere share intuitions of justice about the relative blameworthiness of serious wrongdoing. The four studies that follow are representative.[146]

Michael O'Connell and Anthony Whelan surveyed 623 individuals in the greater Dublin, Ireland area, asking subjects to rate ten offense descriptions on a 1 to 11 scale of seriousness.[147] Clear ordering of offenses emerged, suggesting textured intuitions regarding seriousness and deserved punishment.[148] A comparison of the Irish data with a British sample from a decade earlier found that, "Irish perceptions of crime . . . have much in common with those in other jurisdictions,"[149] especially with regard to more serious crimes.[150]

Following in the tradition of Sellin and Wolfgang, Marlene Hsu administered a survey to six hundred Chinese persons in Taiwan, including students, police officers, and judges.[151] Sub-

146. Joseph E. Scott and Fahad Al-Thakeb compared answers regarding what individuals deemed the "proper punishment" for twenty-four specific crimes among subjects in Denmark, Finland, Kuwait, the Netherlands, Norway, Sweden, and the United States. Joseph E. Scott & Fahad Al-Thakeb, *The Public's Perceptions of Crime: A Comparative Analysis of Scandinavia, Western Europe, the Middle East and the United States*, *in* CONTEMPORARY CORRECTIONS: SOCIAL CONTROL AND CONFLICTS 78, 79 (C. Ronald Huff ed., 1977). Eleven punishments—from no penalty to execution—were available. *Id*. at 80. The researchers found that "[a] great deal of similarity exists among countries as to the rank ordered seriousness" for the subset of offenses involving physical harm and theft. *Id*. at 83. Only slight differences existed. In general, violent offenses, such as murder, rape, and assault, were regarded as more serious than property crimes. *Id*. However, in Kuwait, burglary was rated more serious than assault. *Id*. at 85. Also in Kuwait, drug offenses such as selling heroin or marijuana were also regarded as more serious than offenses such as rape and robbery. *Id*. at 86. Statistical tests were not provided, so it is not possible to know which differences are significant. *Id*. The authors also found "an apparent overall similarity among countries in the perceived seriousness of white-collar offenses." *Id*. They conclude with respect to their international comparisons that, with the exception of Kuwait, "levels of moral indignation for all offenses are much more similar than they are different." *Id*. at 87.

147. O'Connell & Whelan, *supra* note 141, at 302–04. One thousand names were randomly selected from the 1992 electoral register for the greater area of Dublin, and 623 persons responded. *Id*. at 302–03. An example of a tested scenario includes: "The offender attacks a victim with a knife and the victim dies." *Id*. at 304.

148. *Id*. at 310 tbl.2.

149. *Id*. at 316.

150. *Id*. at 310–11.

151. Marlene Hsu, *Cultural and Sexual Differences on the Judgment of Criminal Offenses: A Replication Study of the Measurement of Delinquency*, 64

jects were asked for seriousness judgments of fourteen offenses—which were the fourteen index offenses of Sellin and Wolfgang translated into Chinese—on an eleven-point scale.[152] Hsu found broad agreement across the three Taiwanese samples (students, police, and judges); correlation coefficients were all above 0.90, and similar ordinal judgments in the relative ranking of the fourteen offenses between the Chinese and American samples, with a coefficient of 0.84—0.95 among male subjects.[153] The Sellin and Wolfgang study has been replicated in many other studies in other countries, reaching similar conclusions about shared intuitions of justice.[154]

J. CRIM. L. & CRIMINOLOGY 348, 350 (1973).

152. Some of the offenses were changed slightly to better fit Chinese culture. *Id*. at 349. For example, the stealing of an automobile was changed to stealing a motorcycle because cars were much more rare in China, and as a result, subjects might consider auto theft a more serious offense. *Id*.

153. *Id*. at 351 tbl.2. As noted throughout this Article, characterizing coefficients is subjective. However, these values are very reasonably characterized as "large," particularly given that these are cross-cultural comparisons. *Id*. at 350 tbl.1.

154. In another replication of Sellin and Wolfgang, André Normandeau surveyed 232 introductory sociology students at the University of Montreal—177 males and 55 females. André Normandeau, *The Measurement of Delinquency in Montreal*, 57 J. CRIM. L., CRIMINOLOGY & POLICE SCI. 172, 173 (1966). Subjects were asked to rate fifteen offense descriptions, such as: "An offender forces a female to submit to sexual intercourse. No other physical injury is inflicted." *Id*. at 173 n.7. Subjects rated such crimes on an eleven-point scale. *Id*. Normandeau used Sellin and Wolfgang's fourteen index offenses, but added one item to the list. *Id*. When compared to Sellin and Wolfgang's findings, the results showed a substantial correlation—above 0.9—which led the researchers to conclude that, "there is a large amount of agreement about the numerical scoring of seriousness of offenses between [American and Canadian subjects]." *Id*. at 174. Note that agreement regarding ordering was more substantial than absolute judgments.

In yet another replication of Sellin and Wolfgang, Angel Velez-Diaz and Edwin L. Megargee surveyed 175 individuals in San Juan, Puerto Rico—eighty-three "offenders" and ninety-two "non-offenders." *See* Angel Velez-Diaz & Edwin L. Megargee, *An Investigation of Difference in Value Judgments Between Youthful Offenders and Non-Offenders in Puerto Rico*, 61 J. CRIM. L., CRIMINOLOGY, & POLICE SCI. 549, 550 (1970). "Offenders" consisted of inmates of the Institute of Youthful Offenders. *Id*. The "non-offenders" were drawn from a vocational school close to the Institute of Youthful Offenders. *Id*. The non-offenders were judged to be quite similar to the offenders in most characteristics except, of course, their lack of criminal record. *Id*. Subjects were asked to rate on an eleven-point scale a list of offense descriptions which were the same as those used by Sellin and Wolfgang. *Id*. at 551. The offenses included twenty-one standard offenses and twenty additional offenses, all being drawn from Sellin and Wolfgang's original 141 offenses. *Id*. at 550. They found reasonably high agreement with the Sellin and Wolfgang study when compared to "lower class Puerto Rican offenders and non-offenders," reporting cor-

Graeme Newman sampled 2360 individuals from a number of different cultures—India (512), Indonesia (500), Iran (479), Italy (200), United States (169), and Yugoslavia (500)—chosen for their supposed important cultural differences.[155] Subjects were asked to rate serious offenses, described very briefly and with a minimum of surrounding circumstances, on a twelve-point scale.[156] Newman reports that, "If one were to order the acts according to the proportions of each country sample criminalizing them, one would find a general consensus across all countries as to the extent that all acts should be tolerated."[157] Newman also reports that, "At the general level of analysis, it is apparent that there was considerable agreement as to the amount of official punishment appropriate to each act"[158] and that looking at relative rankings indicates "general agreement in ranks across all countries."[159]

Cross-cultural agreement exists on issues beyond just ranking of harm seriousness. David M. Bersoff and Joan G. Miller presented 180 adults and children aged eight to ten years old in the United States and India with vignettes in which an offender harmed another person or committed a viola-

relations around 0.7 for pairs of samples. *Id*. at 553. Four correlations were drawn: between Puerto Rican offenders and Pennsylvania students, between Puerto Rican offenders and Pennsylvania police, between Pennsylvania non-offenders and students, and between Pennsylvania non-offenders and police. *Id*. at 552 tbl.3. They also performed a test using "Kendall's coefficient of concordance, W," which measures the extent to which there is agreement in how sets of items are ordered. *Id*. at 553. If there was perfect agreement in how all the groups surveyed assigned offenses to the eleven-point scale, a value of 1.0 would be obtained, and a value of 0.0 would indicate no agreement whatsoever. *See infra* text accompanying note 180. A value of 0.8 was obtained, indicating very substantial agreement on relative ordering. Velez-Diaz & Magargee, *supra*, at 553. There was no significant difference in ratings between offenders and non-offenders. *Id*.

155. GRAEME NEWMAN, COMPARATIVE DEVIANCE: PERCEPTION AND LAW IN SIX CULTURES 70, 110–11 tbl.3 (1976).

156. *Id*. at 60. The range included zero, unlike Sellin and Wolfgang's classic work. *Id*.

157. *Id*. at 115.

158. *Id*. at 140.

159. *Id*. at 141, 142–43 tbl.12; *see also id*. at 135–48 (discussing differences in views regarding how particular acts should be controlled or punished). People from different cultures might share the intuition that an act is wrong, and might even agree on an act's relative seriousness, but might differ in how punishment should be imposed—whether by the state, family, or some other source. *Id*. This discussion highlights the importance of assessing intuitions regarding seriousness as distinct from preferred punishments meted out by the state. While the former might be correlated strongly with the latter in some contexts, the correlation will be weaker in others.

tion of property rights, such as theft, and where the circumstances of the offense varied, being either accidental or intentional.[160] The accidental offenses were similarly exculpated across cultures and age ranges: 91% and 100% of the time across samples.[161] This finding indicates broad cross-cultural support for intent as a mediating factor.[162] Other studies reach similar conclusions.[163]

These studies represent only a fraction of the evidence regarding cross-cultural similarities. Jonathan Haidt and his colleagues, in their discussion of the variability in what cultures consider immoral, suggested: "Harm, broadly construed to include psychological harm, injustice, and violation of rights, may be important in the morality of all cultures."[164] Such conclusions obviously resonate closely with our thesis.[165]

---

160. David M. Bersoff & Joan G. Miller, *Culture, Context, and the Development of Moral Accountability Judgments*, 29 DEVELOPMENTAL PSYCHOL. 664, 666–67 (1993).

161. *Id*. at 671. Emotional duress or immaturity of the agent showed some developmental and substantial cross-cultural variability. *Id*.

162. *See* Cecilia Wainryb et al., *Tolerance and Intolerance: Children's and Adolescents' Judgments of Dissenting Beliefs, Speech, Persons, and Conduct*, 69 CHILD DEV. 1541, 1553 (1998) (reporting that "children and adolescents were relatively more accepting of people engaged in potentially harmful or unfair practices who held dissenting informational beliefs than of those who held dissenting moral beliefs" and that subjects "reasoned that persons holding dissenting informational beliefs, although misinformed, were well-intentioned").

163. Similarly, Joseph Sanders and V. Lee Hamilton compared preference for punishment among subjects in Japan, Russia, and the United States for a crime—robbery with accidental death—as opposed to a situation in which a mother strikes her young child. Joseph Sanders & V. Lee Hamilton, *Legal Cultures and Punishment Repertoires in Japan, Russia, and the United States*, 26 LAW & SOC'Y REV. 117, 122–29 (1992). In the former case, subjects in all three cultures expressed universal agreement that the perpetrator should be punished—99%, 99%, and 96%, respectively; for the latter case, there was much less agreement—56%, 21%, and 26%, respectively. *Id*. at 127. The directional effect of mediating factors—mental state, consequences of the criminal's actions, and the perpetrator's history—were all the same, though not always statistically significant. *Id*. at 131.

164. Jonathan Haidt et al., *Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?*, 65 J. PERSONALITY & SOC. PSYCHOL. 613, 613 (1993).

165. There are indeed acts that are considered "wrong" which vary greatly from culture to culture in some domains. The fact that people in one culture think that failing to keep a promise to visit one's mother's grave weekly is something worthy of punishment might seem odd to Westerners, and might seem to indicate vast uncharted depths of moral variability. However, considering acts outside of the deeper conceptual structure risks missing the point. Central to our discussion is that even though different acts are "moralized" from one culture to the next, what makes an act immoral is the concurrent belief that those who perform the act should be punished, and this intuition is

III. TESTING THE LIMITS OF AGREEMENT

As reviewed in the previous Part, earlier studies established the existence of shared intuitions of justice. We sought to build upon that work by investigating its limits. Just how nuanced could these intuitions be? Just how far could one push laypersons in making distinctions among cases with how wide a range of factors?

Our first task was to design a method that minimized testing factors that obscure the extent of agreement. For example, we speculated that both the use of factual descriptions rather than abstract offense labels and the use of full rather than skeletal fact patterns would reduce the problem of inadvertently creating cases that different subjects would perceive differently. We also used a card sorting procedure with scenario headings that might allow an increase in the number and complexity of the distinctions that we asked subjects to make.[166]

We sought to find the outer limits to the nuance and variety of scenarios on which subjects would express consensus. Could we get subjects to distinguish two dozen scenarios on a single continuum of punishment—making quite subtle differences between each adjacent pair—yet still get a high level of agreement in their rank ordering? Could we get subjects to make fine distinctions, and agree in their judgments, when they were judging not just cases that differed in tangible harms, like physical injury, but also differed in intangible harms, such as the extent of intrusion of privacy, or differed in mitigating and excusing conditions, or differed in the vulnerability of the victim?[167]

We used the same test instrument in two different sets of studies. In the first set, reported here as Studies 1 and 3, the test instrument was administered in person to subjects who rank ordered twenty-four cards on a large table.[168] In the second set of studies, identified here as Studies 2 and 4, subjects performed the task over the Internet, using a software program in which they "dragged" each scenario to its appropriate place among the other scenarios. The use of the Internet introduced

widely shared within and between groups. Beneath cultural variability lies the universal intuition that acts that are wrong should be punished. *Id*. The specific acts to which this pertains vary dramatically. *Id*.

166. The study's directions to subjects are reproduced in *infra* Appendix A.

167. See the description of the differences among scenarios summarized in *infra* Appendix B.

168. *See infra* Appendix A.

certain complications in assuring that subjects understood and were motivated to undertake the relatively demanding tasks, but it also allowed for the collection of data from a greater number of subjects with a wider range of demographic characteristics.

To signal our results, we found that, despite the dramatically greater difficulty of this task over those of previous studies, the levels of agreement in rank ordering were astonishingly high. In other words, we failed to find the limits of shared intuitions of justice for core wrongdoing.

A. EXTENT OF AGREEMENT: STUDY 1

1. Method

Participants were given twenty-four short scenarios, each on a separate card describing an event during which "John" engages in conduct that may be a criminal offense.[169] Participants were asked to rank-order the cards on a table to reflect amount of punishment, if any, that John deserves.[170] The scenarios included such offenses as theft by taking, theft by fraud, property destruction, assault, burglary, robbery, kidnapping, rape, negligent homicide, manslaughter, murder, and torture, in a variety of situations, including self-defense, provocation, duress, mistake, and mental illness.[171] The kinds of offenses in the scenarios represent 94.9% of the offenses committed in the United States.[172] When the participant completed the rank ordering, he or she was asked to reconsider each pair of scenarios to confirm that each pair was ordered as wished.[173] This task typically took participants between thirty and forty minutes. Each

169. The text of the twenty-four scenarios is reproduced in *infra* Appendix A.

170. For the directions to participants and administrators, see the end of *infra* Appendix A.

171. *See infra* Appendix A.

172. The offenses in the scenarios, which are the most common offenses committed in the United States, include: sexual assault, 0.8% of all offenses; robbery, 2.5%; assault, 19.0%; household burglary, 14.0%; and theft 58.6%. BUREAU OF JUSTICE STATISTICS, CRIMINAL VICTIMIZATION IN THE UNITED STATES, 2003 STATISTICAL TABLES 14 tbl.1 (2005).

173. The rank ordering task of so many scenarios is quite challenging for many people because it requires a concentration and perseverance beyond that which some people typically are called upon to do in their daily lives. The second run-through was useful for some participants in giving them a chance to review their work.

participant also filled out a form that collected demographic information.[174]

2. Results

Table 1 orders the scenarios in the modal order produced by the subject pool as a group. This is also the predicted rank order.[175] The first four scenarios were most commonly rated as deserving no liability or punishment. Scenario 5 was a borderline case. About a quarter of the subjects thought there should not be punishment, while the remainder thought there should be, but most subjects agreed that it should be ranked as more aggravated than the four no punishment scenarios and less aggravated than the other nineteen scenarios. Mean ranks in ordinal data obviously have limited usefulness but are provided here simply to give a heuristic sense of the distribution for each scenario.[176]

174. For a summary of the demographics of the sample, see *infra* note 183.

175. The reasons underlying the predictions are set out in *infra* Appendix B.

176. For example, the mean value of 23.9 for Scenario 24 affords the inference that nearly every subject ranked this scenario last.

Table 1: Summary Data from Study 1

Scenario	Mode Rank	% Assigning "No Liability"	Mean Rank
S1 (defending)	no punish	91	*
S2 (coercion)	no punish	92	*
S3 (umbrella)	no punish	92	*
S4 (hallucination)	no punish	92	*
S5 (pies)	5	27	3.9
S6 (T-shirt)	6	8	6.3
S7 (short change)	7	9	6.8
S8 (radio)	8	5	7.6
S9 (drill)	9	3	8.8
S10 (microwave)	10	2	9.9
S11 (TV)	11	2	10.4
S12 (slap)	12	0	12.1
S13 (head-butt)	13	2	13.0
S14 (stitches)	14	0	14.1
S15 (necklace)	15	0	14.8
S16 (robbery)	16	0	15.6
S17 (clubbing)	17	0	16.8
S18 (pit bulls)	18	0	18.2
S19 (infant)	19	3	18.6
S20 (stabbing)	20	0	19.7
S21 (ambush)	21	0	20.9
S22 (abduction)	22	0	21.7
S23 (burning)	23	0	23.0
S24 (ransom)	24	0	23.9
N = 24			

* The means for these four scenarios are 0.6, 0.3, 1.4, 0.3, respectively, where a response of "no punishment" is given a value of 0.

The task asked of the subjects was quite complex. Ranking each scenario required a comparison to each of the other scenarios. We expected an unavoidable amount of "noise" in the data to reflect the fact that many subjects would have difficulty sustaining the required level of concentration for the length of time required for the task. Thus, it was quite surprising that the results nonetheless were rather clear and dramatic.

For the first four scenarios, there was overwhelming consensus that "no punishment" was appropriate. Ninety-two percent of the time subjects agreed that no punishment was deserved. The sixty-four subjects each judged these four scenarios—a total of 256 judgments—and in 235 instances concluded that no punishment was deserved.

Table 2 lists the remaining twenty scenarios in increasing order based on the subject group's modal ranking. Each cell shows the frequencies with which subjects deviated from the group's modal ranking of each pair of scenarios. In other words,

it counts the "deviations" from the group judgment. For example, the intersection of S7 (short change scenario) and S9 (drill theft scenario) indicates that nine subjects ranked S9 as less serious than S7, deviating from the more common view that S9 (drill theft) was more serious than the S7 (short change).

Table 2: Deviations from Modal Ranks in Study 1

	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24
S5	0	4	5	3	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
S6		0	23	12	5	2	2	3	2	1	0	1	0	0	2	0	0	0	0	0
S7			0	17	9	7	9	3	3	2	1	0	1	0	1	0	0	0	0	0
S8				0	12	5	5	3	3	2	1	1	1	0	1	0	0	0	0	0
S9					0	10	7	4	3	2	1	2	1	0	2	0	0	0	0	0
S10						0	13	6	3	2	1	1	1	0	2	0	0	0	0	0
S11							0	6	3	2	1	1	1	0	2	0	0	0	0	0
S12								0	9	6	7	2	0	2	4	0	0	0	0	0
S13									0	9	11	4	1	2	4	0	0	0	0	0
S14										0	19	9	2	2	4	0	0	0	0	0
S15											0	20	5	3	3	0	0	0	0	0
S16												0	6	4	4	0	0	0	0	0
S17													0	4	5	1	0	0	0	0
S18														0	19	5	2	1	0	0
S19															0	19	10	9	3	1
S20																0	4	3	1	0
S21																	0	10	1	0
S22																		0	2	0
S23																			0	5
S24																				0

N = 64

NB: Shading indicates instances where adjacent scenarios were "flipped" by subjects in their rankings.

Each of the sixty-four subjects could have deviated from the group modal ranking as to each of the 190 pairwise judgments—the number of cells on the table[177]—for a total number of 12,160 possible deviations from modal pairwise ranking.[178] Note that if a subject ranks a scenario different from the modal ranking, that choice may produce many pairwise "deviations." For example, a subject who agreed with the modal ranking in every respect, except that he ranked S11 much higher, as more

177. Where n is the number of scenarios, this is $n(n\text{-}1)/2$.

178. Of course, as a practical matter, this value overstates the total number of possible deviations because we rank the scenarios in order based on the observed modal rankings. Consider Scenario 5. Because the most frequent ranking of this scenario was the least serious of the twenty punishment scenarios, we know that at least those subjects who ranked Scenario 5 as the least serious could not, in principle, produce a deviation between Scenario 5 and the remaining scenarios, effectively limiting the total number of possible deviations.

serious than S19, would, by his ranking of that one scenario, produce eight deviations. That is, he would have deviated from the judgments of the other subjects in eight different pairwise judgments, S11 versus S12, S11 versus S13, etc.

The subjects had a total of 480 deviations, meaning that *the subjects reversed the modal pairwise ranking only 4% of the time*. This means that subjects agreed with the modal pairwise ranking 96% of the time, as compared to the approximately 50% agreement rate that a random ranking of pairs would give.[179]

The most common deviation, as one might guess, was for a subject to "flip" the ranking of two scenarios that were adjacent in the group's modal ranking—for example, a subject ordering the scenarios as S6, S8, S7, S9, "flipping" scenarios S7 and S8. Of the total 480 deviations, 211 were such "adjacent flip" deviations. *If these simple "flips" of adjacent scenarios are excluded, the percentage of pairwise rankings that deviate from the mode is only 2%.*

More sophisticated statistical methods of analysis are available, most importantly Kendall's coefficient of concordance (Kendall's W), which measures agreement (concordance) among different sets of rank orderings of the same set.[180] This coefficient measures the concordance shown in the entire ranking, not just in a set of pairwise comparisons. A coefficient of 1.0 indicates perfect agreement, a 0.0 indicates no agreement.[181] *For the twenty-four scenarios in the study, the Kendall's W is 0.95* (with $p < 0.001$).[182] This is an astounding level of agreement.

The participants in the study were a diverse group in some ways but not others.[183] The group was not an accurate propor-

179. A random set of responses would have roughly 50% of agreement with the modal pairwise ranking. Thus, the relevant point of comparison here is the amount of agreement over 50%, not over 0%.

180. *See* JERROLD H. ZAR, BIOSTATISTICAL ANALYSIS 390 (3d ed. 1996).

181. *See id*.

182. N = 64; Chi-square = 1397.9; df = 23. The p value indicates how unlikely it is that this result would occur due to chance if people were responding at random.

183. Here is what the participants look like as a group:

Gender: male 36%, female 64%

Marital status: single 23%, married 58%, divorced 9%, widowed 8%

Have children?: yes 70%, no 30%

Race: white 91%, nonwhite 9%

Education: some college 5%, two-year college degree 2%, four-year college degree 38%, masters degree 39%, doctorate/professional degree 17%

tional representation of the society generally, but this fact was of little consequence given that the claim being investigated did not relate to representativeness. With so little disagreement among the group, combined with the subject pool size of sixty-four, it would be difficult if not impossible to find variation in results according to demographic variables.

3. Discussion

Preliminarily, note that the present study confirms the rather basic conclusion of Part II.A that people do indeed share an intuition that serious wrongdoing should be punished. Each subject in the study could have judged "no liability" for any or all of the twenty-four scenarios presented.[184] Instead, subjects typically chose "no liability" only for the first four scenarios, which were designed to elicit a judgment of full exculpation.[185] For every other scenario, subjects typically expressed an intuition that at least some punishment was deserved.

As to shared intuitions regarding the relative seriousness of wrongdoing, discussed in Part II.B, the amount of agreement shown in the study—96% of all pairwise judgments, and a Kendall's W of 0.95—represents an extraordinary result. To appreciate its significance, consider some points of comparison. For example, American men who ranked the attractiveness of women with different waist-to-hip ratios (WHRs) generally agreed in their rankings, showing a Kendall's W of 0.54.[186] Readers of a travel magazine asked to rank eight travel destinations—such as Israel, New York, and Canada—according to the risk of terror generally agreed in their rankings, with a Kendall's W of 0.52.[187] On the other hand, when economists

Age: 20 or under 2%, 21–30 11%, 31–40 5%, 41–50 11%, 51–60 22%, 61 or over 47%

Political orientation: very liberal 5%, liberal 27%, somewhat liberal 28%, center 17%, somewhat conservative 14%, conservative 9%

Religion: Jewish 11%, Protestant 77%, other Christian 9%, other 2%

Level of religious activity: very active 47%, active 25%, somewhat active 19%, not active 8%

Income (annual in thousands): 20–40 5%, 40–60 16%, 60–80 6%, 80–100 19%, more than 100 31%

184. *See infra* Appendix A (reporting instructions to the participants).

185. *See supra* Table 1. In these four scenarios, the act that caused harm was either unintentional (a mistake or a hallucination), coerced, or self-defense. For more on the scenarios' analysis, see *infra* Appendix B.

186. Frank Marlowe & Adam Westman, *Preferred Waist-to-Hip Ratio and Ecology*, 30 PERSONALITY & INDIVIDUAL DIFFERENCES 481, 483 (2001).

187. Baruch Fischhoff et al., *Travel Risks in a Time of Terror: Judgments*

were asked to rank the top twenty economic journals according to quality, the Kendall's W was 0.095.[188]

Very high Kendall's W scores typically are found only in tasks that seem quite obvious, as when subjects are asked to rank order images according to their brightness, yielding a Kendall's W of 0.95,[189] or, to take another example, when subjects are asked to rank the following six drawings of faces according to how much pain they show, giving a Kendall's W of 0.97.[190]

Figure 1: Faces Pain Scale

One may wonder how this amount of agreement could exist when judging such a complex matter as relative blameworthiness. It would seem to call for a matter of complex and subjective judgment—like judging the relative talent of twenty movie actors or athletes, the beauty of twenty flower arrangements or paintings, the awkwardness of twenty embarrassing situations,

and Choices, 24 RISK ANALYSIS 1301, 1303 (2004).

188. Kostas Axarloglou & Vasilis Theoharakis, *Diversity in Economics: An Analysis of Journal Quality Perceptions*, 1 J. EUR. ECON. ASS'N 1402, 1421 (2003) ("These results unveil significant diversity in the journal quality perceptions among groups of economists despite the fact that our sample focused on AEA members. To test the robustness of this claim, using Kendall's W we examined the correlation in journal quality perceptions between any two randomly selected economists in our sample. We found Kendall's W for the top ten journals in our rankings to be 0.396, which demonstrates a relatively low level of agreement among economists. Once we extended this exercise to the top 20 journals in our rankings, Kendall's W dropped to only 0.095.").

189. Charles M.M. de Weert & Noud A.W.H. van Kruysbergen, *Assimilation: Central and Peripheral Effects*, 26 PERCEPTION 1217, 1221 (1997). The images are made up of black and white splotches. *Id*. at 1222–24. The more white splotches, the brighter the image appears. *Id*. at 1224.

190. Keela A. Herr et al., *Evaluation of the Faces Pain Scale for Use with the Elderly*, 14 CLINICAL J. PAIN 29, 29 (1998) ("Rank ordering tasks for the individual faces demonstrated near-perfect agreement between the actual expected ranking and the ranking produced by the subjects (Kendall's W = .97)."). The faces are used to collect diagnostic information from patients who are unable to communicate orally. *See* Pediatric Pain Sourcebook, Submission and Review Form, http://painsourcebook.ca/pdfs/pps92.pdf (last visited Apr. 21, 2007).

or the humorousness of twenty comedians or clowns. How is it that, in contrast, people can agree so much with one another when asked to judge the relative punishment deserved in twenty-four crime scenarios, a task that would seem to involve equally complex matters of judgment? The existence of this level of agreement presents an intriguing puzzle in itself.[191]

B. EXTENT OF AGREEMENT: STUDY 2

In a follow-up study, 246 subjects rank-ordered the same twenty-four scenarios using a computer-based program over the Internet.[192] Despite the potential for a large increase in the amount of "noise," the results were essentially the same.[193]

1. Method

In the computer-based test, each Internet participant would "drag" each scenario to that location in a "stack" of scenarios that reflected the proper rank of the scenario with regard to the amount of punishment deserved. As in the paper-based administration, subjects were free to identify any or all scenarios as deserving of no punishment.

Study 2 was conducted through the National Science Foundation's Time-Sharing Experiments for the Social Sciences (TESS) program.[194] This method resulted in some important differences with respect to the study described above. First, subjects were recruited using a random sampling procedure which draws broadly from across the United States. As can be demonstrated by the demographic data presented below, subjects came from a variety of socioeconomic, religious, and racial backgrounds.[195] Participants were presented with a web-based

191. We attempt a preliminary answer to that puzzle in Paul H. Robinson, Robert Kurzban & Owen Jones, *The Origins of Shared Intuitions of Justice*, 60 VAND. L. REV. (forthcoming 2007).

192. The responses of four of the 250 subjects were excluded from the dataset because they did not follow the task instructions or because their responses were essentially random. With these four cases included, the Kendall's W for the full 250 subjects is 0.86.

193. *See infra* Part III.B.3.

194. Data collected by Time-Sharing Experiments for the Social Sciences (TESS), National Science Foundation Grant 0094964, Diana C. Mutz, Principal Investigator. Award Abstract #0094964, http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0094964 (last visited Apr. 21, 2007).

195. *See infra* note 199. Polimetrix Information implemented and completed the present data collection. Information about their quota sampling technique is available online at Polimetrix: The Science of Political Measure-

interface that asked them to perform the same task described above. Of course, because no experimenter was present at the time the task was performed, there is substantially increased potential for "noise," as there is no way to directly observe whether subjects were attending to the task. The motivation of subjects when they sit in front of a computer terminal may be considerably less than when they have face-to-face contact with a human researcher. For a task that demands as much time and concentration as the rank ordering of twenty-four scenarios, the lack of personal motivation, as well as the lack of quality control that comes from personal observation, meant that the resulting data could be quite "noisy."

There were indications that these concerns were well-founded. Many of the subjects gave results that would seem inexplicable to the vast majority of their fellow participants. For example, one subject ranked the offender who shortchanged the customer of twenty dollars in Scenario 7 as deserving of more punishment than the offender who did the vicious stabbing in Scenario 20, and ranked the T-shirt theft in Scenario 6 as deserving of more punishment than the ransom, rape, torture, and strangling in Scenario 24. Is it likely that this subject understood the instructions and was taking the task seriously? Another subject ranked the person who stole the camera because he was coerced to do so by a threat to his child, in Scenario 2—judged as deserving no punishment by most subjects—as deserving of more punishment than the brutal abduction and killing in Scenario 22. Another subject thought that taking the clock radio from the car, in Scenario 8, deserved more punishment than the beating of the victim with a club during a robbery in Scenario 17, and that the head-butt causing a gash requiring stitches, in Scenario 13, deserved less punishment than taking pies from the all-you-can-eat buffet in Scenario 5. None of these subjects' responses was excluded from the dataset.

A final difference with the previous study is that a slightly different set of demographic questions was administered at the conclusion of the study. These demographic items in addition to a few other items are routinely gathered by Polimetrix.

ment, http://www.polimetrix.com/services_sample.html (last visited Apr. 21, 2007).

2. Results

Table 3 presents the scenarios in the modal order of the participants as a group. As in paper-based Study 1, the first four scenarios were typically ranked as not deserving punishment. Scenario 5 was seen as a borderline case for which subjects disagreed about whether punishment was deserved but agreed that it should be ranked as more aggravated than the four "no punishment" scenarios but less aggravated than the other nineteen scenarios. As with Table 1, mean ranks in ordinal data obviously have limited usefulness but are provided here simply to give a heuristic sense of the distribution for each scenario.

Table 3: Summary Data from Study 2

Scenario	Modal Rank	% Assigning "No Liability"	Mean Rank
S1 (defending)	no punish	82	*
S2 (coercion)	no punish	75	*
S3 (umbrella)	no punish	87	*
S4 (hallucination)	no punish	71	*
S5 (pies)	5	8	5.4
S7 (short change)	6	6	7.1
S6 (T-shirt)	7	1	7.4
S8 (radio)	8	0	8.4
S9 (drill)	9	0	9.2
S10 (microwave)	10	0	10.2
S11 (TV)	11	0	10.6
S12 (slap)	12	0	11.6
S13 (head-butt)	13	2	12.2
S14 (stitches)	14	0	13.7
S15 (necklace)	15	0	13.6
S16 (robbery)	16	0	15.3
S17 (clubbing)	17	1	16.7
S18 (pit bulls)	18	0	17.8
S19 (infant)	19	1	19.0
S20 (stabbing)	20	0	20.1
S21 (ambush)	21	0	20.9
S22 (abduction)	22	0	21.4
S23 (burning)	23	0	22.2
S24 (ransom)	24	0	23.5
N = 246			

* The means for these four scenarios are 1.2, 2.3, 0.3, 1.7, respectively, when a response of "no punishment" is given a value of 0. Mean ranks in ordinal data obviously have limited usefulness but are provided here simply to give a heuristic sense of the distribution for each scenario, and to establish the modal relative relation among the scenarios for the layout of Table 4.

As in the previous study, the first four scenarios were judged to merit "no punishment" by an overwhelming proportion of subjects: 82%, 75%, 87%, and 71% for the four scenarios respectively.[196] Table 4 lists the remaining twenty scenarios in increasing order based on the subject group's modal ranking. As with Table 2, each cell shows the frequencies with which subjects deviated from the group's modal ranking of each pair of scenarios.

Table 4: Deviations from Modal Ranks in Study 2

	S5	S7	S6	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24
S5	0	59	54	30	26	19	13	12	15	7	12	4	2	1	4	0	0	0	0	1
S7		0	117	82	54	47	48	27	27	17	16	8	4	2	5	1	1	1	1	1
S6			0	79	61	43	46	28	31	15	14	8	6	3	5	1	1	1	1	2
S8				0	78	52	49	43	42	18	27	10	7	4	7	2	0	1	2	2
S9					0	74	62	53	52	32	32	12	8	3	5	1	0	0	0	1
S10						0	100	66	61	36	40	13	9	5	6	3	1	2	2	3
S11							0	75	67	43	41	18	8	6	4	2	1	1	1	2
S12								0	84	43	62	18	9	7	8	1	1	1	1	1
S13									0	67	87	36	18	9	12	2	1	1	2	2
S14										0	117	65	23	14	19	2	0	0	1	1
S15											0	64	36	15	19	1	2	2	1	1
S16												0	57	33	23	3	0	1	2	2
S17													0	62	43	13	4	3	6	5
S18														0	64	30	20	16	6	4
S19															0	90	61	52	37	12
S20																0	76	54	35	10
S21																	0	85	58	18
S22																		0	66	19
S23																			0	33
S24																				0

N = 64

NB: Shading indicates instances where adjacent scenarios were "flipped" by subjects in their rankings.

Each of the 246 subjects could have deviated from the group modal ranking as to each of the 190 pairwise judgments—the number of cells on the table—for a total number of 46,740 possible deviations from modal pairwise ranking. Recall that if a subject ranks a single scenario different from the modal ranking, that choice may produce many pairwise "deviations." The subjects had a total of 4317 deviations, meaning that *the subjects reversed the modal pairwise ranking only 9.2% of the time. That is, they agreed with the pairwise modal rank*

196. Overall, 79% of the time subjects agreed that no punishment was deserved in these four scenarios. The 246 subjects each judged these four scenarios—a total of 984 judgments—and in 774 instances concluded that no punishment was deserved.

91.8% of the time, as compared to the roughly 50% agreement rate that a random ranking of pairs would give.[197]

The most common deviation, as one might guess, was for a subject to "flip" the ranking of two scenarios that were adjacent in the group's modal ranking—for example, a subject rank ordering the scenarios as S6, S8, S7, S9, "flipping" scenarios S7 and S8. There were 1447 such "adjacent flip" deviations, accounting for about a third of all deviations. *If these simple "flips" of adjacent scenarios are excluded, the percentage of pairwise rankings that deviate from the mode is only 6.1%.*

As in paper-based Study 1, we calculated a Kendall's coefficient of concordance for the subjects rankings. Recall that a coefficient of 1.0 indicates perfect agreement, a 0.0 indicates no agreement.[198] For the 246 subjects rank-ordering the twenty-four scenarios, *the Kendall's W is 0.88* (with $p < 0.001$), again, a surprisingly high level of agreement.

The participants in the computer-based study were quite diverse.[199] Table 5 shows the modes for some of the more important demographic groups.

---

197. Remember, if subjects were responding randomly to the pairwise comparisons, they would agree 50% of the time, so the point of comparison here is the amount over 50%, not over 0%.

198. *See* ZAR, *supra* note 180, at 390.

199. The participant group looks like this:

Gender: male 50%, female 50%

Race: white 70%, black 4%, Hispanic 6%, Native American 1%, mixed 5%, other 4%

Age: 20 or under 6%, 21–30 15%, 31–40 21%, 41–50 18%, 51–60 17%, 61 or over 23%

Marital status: single 23%, married 61%, divorced 8%, widowed 4%, domestic partnership 4%, separated 2%

Have children?: yes 64%, no 36%

Education: no high school degree 4%, high school graduate 18%, some college 47%, two-year college degree 7%, four-year college degree 14%, graduate degree 10%

Political orientation: very liberal 11%, liberal 20%, moderate 38%, conservative 22%, very conservative 9%

Religion: Jewish 4%, Protestant 33%, Muslim 4%, Catholic 2%, no religion 23%, other 18%

Level of religious activity: very active 22%, active 11%, somewhat active 18%, not active 49%

Income (annual in thousands): less than 20 8%, 20–40 16%, 40–60 18%, 60–80 14%, 80–100 8%, over 100 20%, decline to answer 17%

Registered to vote?: yes 95%, no 5%

Strength of political views: strong Democrat 20%, Democrat 10%, leaning Democrat 8%, independent 18%, leaning Republican 7%, Republican 10%, strong Republican 17%, not sure 9%

Table 5: Modal Rankings Broken Down by Demographics, Study 2

Scenario		All Subjects	Male	Female	Non-White	White	<$60K Income*	>$60K Income*	<2 yr Degree	≥2 yr Degree
N =		246	123	123	53	193	102	103	169	77
S1	(defending)	0†	0	0	0	0	0	0	0	0
S2	(coercion)	0	0	0	0	0	0	0	0	0
S3	(umbrella)	0	0	0	0	0	0	0	0	0
S4	(hallucination)	0	0	0	0	0	0	0	0	0
S5	(pies)	5	5	5	5	5	5	5	5	5
S7	(short change)	6	6	6	7	6	6	6	6	6
S6	(T-shirt)	7	6	7	6	7	7	7	7	7
S8	(radio)	8	8	8	7,8‡	8	8	8	8	8
S9	(drill)	9	9	9	9	9	9	9	9	9
S10	(microwave)	10	10	11	11	10	11	10	11	10
S11	(TV)	11	11	11	8	11	11	11	11	11
S12	(slap)	12	12	12	12	12	12	12	12	12
S13	(head-butt)	13	13	13	13	13	13	13	13	13
S14	(stitches)	14	14	14	14	14	14	14	14	15
S15	(necklace)	15	15	15	15	15	15	15	15	14
S16	(robbery)	16	16	16	16	16	16	16	16	16
S17	(clubbing)	17	17	17	17	17	17	17	17	17
S18	(pit bulls)	18	18	18	18	18	18	18	18	18
S19	(infant)	19	19	19	19	19	19	19	19	19
S20	(stabbing)	20	20	20	20	20	20	20	20	20
S21	(ambush)	21	21	21	21	21	21	21	21	21
S22	(abduction)	22	22	22	22	22	22	22	22	22
S23	(burning)	23	23	23	23	23	23	23	23	23
S24	(ransom)	24	24	24	24	24	24	24	24	24

* Forty-one subjects did not provide income information.
† "No punishment" as the modal response is shown as 0.
‡ The two ranks were a tie, thus both modes are reported.

There appears to be little variation in the modes of scenario rankings according to these demographic variables.[200] Given the high level of agreement, it would be difficult to find variations in rankings as a function of demographic groups.[201]

Libertarian: not at all 60%, a little 24%, somewhat 40%, very 15%, extremely 4%, decline to answer 1%

200. An investigation of the other demographic variables on which data was collected from the subjects—political party, ideology, marital status, whether they have children, religion, level of religious activity, libertarianism—showed a similar lack of any meaningful difference between demographic groups' modal rankings. TESS Demographic Modes Table (on file with the authors).

201. We plan to examine this issue using more sophisticated statistical analyses. *See infra* note 227 and accompanying text.

3. Discussion

The results of computer-based Study 2 reinforce the conclusions of paper-based Study 1. Both studies show an exceptionally high level of agreement among the participants. The result in Study 2 is all the more surprising given the great potential for "noise" in a study where researchers cannot supervise or even observe the participants as they perform the task. The subjects agreed with the modal pairwise ranking 92% of the time and, if the "flips" of adjacent scenarios are excluded, they agreed with the modal pairwise ranking 94% of the time. The Kendall's W coefficient of concordance was 0.88. This result confirms the astounding level of agreement shown by Study 1's Kendall's W of 0.95, in performing what appears to be a rather complex and subjective judgment task.

IV. DISAGREEMENTS ON INTUITIONS OF JUSTICE

Disagreements on intuitions of justice do exist. Some of the apparent disagreements are not real, but the studies reported below suggest that true disagreements do exist for intuitions about wrongdoing outside the core of physical aggression, unconsented-to takings, and deception or deceit in exchanges.

A. APPARENT DISAGREEMENTS AMONG INTUITIONS OF JUSTICE

People obviously do disagree about many things relating to crime and punishment, as the endless public debates make clear. What can be said about the issues on which there is agreement and the issues on which there is not?

Some apparent sources of disagreement are simply misleading. For example, poor testing methods will predictably underestimate the extent of agreement. As noted previously, when a test scenario is written ambiguously so that different test participants perceive it differently, the existence of shared intuitions of justice itself will predict different judgments among the participants.[202] So too, when a case in the headlines has social or political implications, it is common that its relevant facts will be perceived differently by different people. What one makes of the police testimony in the O.J. Simpson case or the Rodney King case may depend upon one's experiences and one's peers' experiences with police officers in daily life. If people draw different conclusions from the testimony,

202. *See supra* Part II.B.1.

they may have different views of the relevant facts of the case, which would predict different views on the liability and punishment deserved.

Another source of apparent disagreement has been mentioned above.[203] While people may agree on the relative blameworthiness of a set of cases, some people may prefer generally harsher punishments than other people.[204] That is, some people may set the most severe end of the punishment continuum noticeably higher than others, which would predict different sentences, even if the people agreed on the relative blameworthiness of different offenders.

There is evidence that attitudes towards the severity of punishment generally vary with some demographics, such as race and socioeconomic status. For example, whites have been found to support the death penalty more than blacks.[205] Similarly, whites are more supportive of "three strikes" laws.[206] Religious variables also have effects on issues such as the death penalty, though their effects are not straightforward.[207] Generally, women have been found to be less punitive than men.[208] Data from Canada and elsewhere indicate that people with less education and lower incomes compared to those with more education and higher income tend to be more punitive in the sense that they report that sentences are "not severe enough,"[209]

203. *See supra* Part II.B.

204. *See supra* Part II.B.

205. Robert L. Young, *Race, Conceptions of Crime and Justice, and Support for the Death Penalty*, 54 SOC. PSYCHOL. Q. 67, 67 (1991).

206. Yesilernis L. Peña et al., *Race and Support for the Criminal Justice System: A Matter of Asymmetry* 29 tbl.1 (Russell Sage Found., Working Paper No. 181, 2002), *available at* http://www.russellsage.org/publications/workingpapers/Race%20and%20Support%20for%20the%20Criminal%20Justice%20System/document.

207. Robert L. Young, *Religious Orientation, Race and Support for the Death Penalty*, 31 J. SCI. STUDY RELIGION 76, 82 (1992) ("Membership in a fundamentalist church and belief in Biblical literalism increased support, while evangelism was associated with reduced support.").

208. Felicia Pratto et al., *The Gender Gap: Differences in Political Attitudes and Social Dominance Orientation*, 36 BRIT. J. SOC. PSYCHOL. 49, 49 (1997) ("Results [from a recent study] replicate previous findings of more male support of conservative ideology, military programmes, and punitive policies . . . .").

209. Carla Cesaroni & Anthony N. Doob, *The Decline in Support for Penal Welfarism: Evidence of Support Among the Elite for Punitive Segregation*, 43 BRIT. J. CRIMINOLOGY 434, 438 (2003).

though the role of education on policies relating to punitiveness is somewhat complex.[210]

The same differences are seen cross-culturally. Average prison sentences vary widely from nation to nation. American offenders were required to serve an average of twenty-nine months after conviction in 1999.[211] In contrast, the average offender in the Netherlands was released after five months,[212] while Columbian offenders were not released until a startling mean of 140 months.[213] Moreover, even within a culture, community attitudes toward punishment severity can vary over time.[214] The effect of these different views on sentencing severity generally can be to obscure agreement on the relative blameworthiness of specific cases.

Another effect that exaggerates the extent of disagreement is the abstraction and politicalization process. When people are asked whether they support or oppose the death penalty, for example, their answers may be different from when the question is posed in terms of a specific set of facts with a specific offender and victim.[215] This phenomenon has direct analogs in other areas of psychology. When persons in need are identified individuals, for which specific and concrete information is provided, people are more likely to help them than when such information is not provided.[216] Stereotypes are more commonly

210. Michael J. Leiber et al., *The Effects of Occupation and Education on Punitive Orientations Among Juvenile Justice Personnel*, 30 J. CRIM. JUST. 303, 303 (2002) (showing that for the sample studied—juvenile justice personnel—"[i]ncreases in education reduced adherence to punishment orientations"). For an interesting discussion and recent data, see Christopher M. Federico & Justin W. Holmes, *Education and the Interface Between Racial Perceptions and Criminal Justice Attitudes*, 26 POL. PSYCHOL. 47 (2005).

211. UNITED NATIONS OFFICE ON DRUGS AND CRIME, SEVENTH UNITED NATIONS SURVEY ON CRIME TRENDS AND THE OPERATIONS OF CRIMINAL JUSTICE SYSTEMS 480 (1998–2000).

212. *Id*. at 308.

213. *Id*. at 66.

214. *See, e.g.*, MICHAEL TONRY, SENTENCING MATTERS 137 (1996) (noting that the average prison sentence for violent offenses in the United States tripled between 1975 and 1989).

215. *See* CRAIG HANEY, DEATH BY DESIGN: CAPITAL PUNISHMENT AS A SOCIAL PSYCHOLOGICAL SYSTEM 76 (2005) (suggesting that people imagine defendants to be more dangerous than the typical defendant when polling questions are phrased in an abstract way).

216. Tehila Kogut & Ilana Ritov, *The "Identified Victim" Effect: An Identified Group, or Just a Single Individual?*, 18 J. BEHAV. DECISION MAKING 157, 163 (2005); *see also* Deborah A. Small & George Loewenstein, *Helping* a *Victim or Helping* the *Victim: Altruism and Identifiability*, 26 J. RISK & UNCERTAINTY

applied to groups—for example, black males—than to specific individuals in the category.[217] In both of these cases, as with the death penalty, the difference between the specific and the abstract changes responses.[218] Hence, when a criminal punishment issue or case takes on a broader public profile, one's view on it often carries baggage that can distort what a person would otherwise see as her intuition of justice.

B. TRUE DISAGREEMENTS AMONG INTUITIONS OF JUSTICE: STUDIES 3 AND 4

Notwithstanding these many false appearances of disagreements among intuitions of justice, it is true that there are punishment-assignment issues on which people do indeed disagree. Two additional studies, using the same participant pools as Studies 1 and 2 above, illustrate the potential for disagreement and may hint at its source.

1. Method: Studies 3 and 4

After completing the paper-based rank ordering of the twenty-four scenarios of Study 1, participants were given an additional twelve scenarios.[219] The methodology was the same as used in Study 1: Each scenario was printed on a separate card, which the participants were asked to rank order on a large table according to the amount of punishment deserved, if any.[220] In addition, participants were asked to show how the punishment deserved in these twelve scenarios compared to that in the first set of twenty-four scenarios, by laying the twelve new cards out on the table next to the original twenty-four cards, using the placement of the card to indicate the

5, 13 (2003) ("In combination, these two studies provide new evidence supporting the existence of an identifiable victim effect.").

217. *See* Michael J. Gill, *Biased Against "Them" More Than "Him": Stereotype Use in Group-Directed and Individual-Directed Judgments*, 21 SOC. COGNITION 321, 321 (2003) ("Stereotypes predicted group-directed social policy judgments but showed little relation to individual-directed impression or social policy judgments.").

218. There are other similar effects. *See, e.g.*, Eric J. Johnson et al., *Framing, Probability Distortions, and Insurance Decisions*, 7 J. RISK & UNCERTAINTY 35, 40 (1993) (noting that, for example, people are more willing to pay for insurance against something that is concrete and specific rather than something stated more abstractly: "The total price [of insurance that the subjects were willing to pay] reported for disease and then accident protection is more than twice that reported for protection for 'any reason.'").

219. The text of the twelve scenarios is reproduced in *infra* Appendix C.

220. *See infra* Appendix A.

proper punishment in relation to the previously ranked scenarios. Participants could place the twelve new scenarios either next to or between the previously ranked twenty-four scenarios.

The twelve scenarios of this second group included offenses of drunk driving, prostitution, marijuana purchase, purchase of alcohol for use by teenagers, bestiality, late-term abortion, cocaine purchase, date rape, third felony offense (jewelry grab), large-scale cocaine selling, and very large-scale cocaine importation and distribution.[221] While a majority of these offenses are committed with much lower frequency than the offenses in Study 1,[222] they are each in their own turn an offense on which there has been public debate.

In the final study, Study 4, the same participants who had ranked the first twenty-four scenarios using the Internet-based procedures in Study 2 now ranked the same additional twelve scenarios used in paper-based Study 3 described above. In the Internet version, however, participants simply ranked the twelve scenarios among themselves, without indicating how they compared to the twenty-four scenarios they had previously ranked in Study 2. This was largely a concession to logistic limitations of the web-based interface.

2. Study 3 Results

Table 6 lists the scenarios in the order in which the subjects as a group ranked them, according to mean rank. Recall that the subjects are using their ranking of the first set of twenty-four scenarios as their "rank calibration," in ranking this second set of twelve scenarios. Thus, the point of reference for understanding the ranks in Table 6 below is the ranks established in Study 1, presented in the second column of Table 3. The lettering of the scenarios indicates the order of seriousness as reflected in the statutory penalties commonly associated with each offense, as illustrated in Appendix D. Scenario A typically carries the lowest statutory penalty, scenario L the highest.

221. *See infra* Appendix C.
222. *See supra* note 172 and accompanying text.

Table 6: Summary Data from Study 3

Scenario	Mean Rank*	Mode Rank (frequency)†	% Assigning "No Liability"
E (bestiality)	5.1	no punish (27)	42.2
B (prostitution)	7.3	no punish (15)	23.4
C (marijuana)	7.4	no punish (14)	21.9
G (cocaine)	10.2	11–12 (9)	9.4
A (drunk crash)	12.0	11–12 (9)	3.1
D (teen alcohol)	12.2	17–18 (10)	4.7
I (3rd theft)	13.6	11–12 (12)	0.0
F (late abortion)	14.0	no punish, 17–18 (8)	12.5
H (unwanted sex)	15.5	17–18 (17)	1.6
K (cocaine dealer)	16.5	17–18 (21)	3.1
L (cocaine importer)	17.4	17–18 (24)	0.0
J (rape)	18.6	17–18 (18)	0.0
N = 64			

* Recall that subjects' ranks in this Study relied upon the rankings they had given in the first set of scenarios, in Study 1. As noted in the context of Table 1, mean ranks in ordinal data obviously have limited usefulness but are provided here simply to give a heuristic sense of the distribution for each scenario and to establish the modal relative relation among the scenarios for the layout of Table 7. For example, the 18.6 mean rank for scenario J suggests that subjects most closely associated this scenario with the seriousness of scenarios 18 and 19 of Study 1.

† A mode of 11–12 on the table indicates that the most common subject response was to rank the scenario as being between the scenarios in the first set that she had ranked as 11th and 12th. The number in parentheses indicates the number of subjects who gave that modal response.

Note the low number of responses that created the modal response for each scenario, suggesting that the responses were quite scattered.

Using this mean ranking to order the scenarios according to their relative seriousness as perceived by the subjects, Table 7 follows this order in setting scenarios along its two axes. The body of the Table shows the frequencies with which subjects' judgments about the ranking of each scenario differed from this order of ranking. The Table contains the same kind of information as shown in Tables 2 and 4 previously: the frequencies with which subjects reversed their ordering from that of the group as a whole.

Table 7: Deviations from Modal Ranks in Study 3

Scenario	E	B	C	G	A	D	I	F	H	K	L	J
E (bestiality)	0	14	15	9	5	6	4	4	3	3	3	0
B (prostitution)		0	19	7	7	6	4	8	4	4	2	0
C (marijuana)			0	2	7	5	10	8	3	0	0	1
G (cocaine)				0	14	12	13	13	8	1	0	0
A (drunk crash)					0	23	20	19	10	7	3	4
D (teen alcohol)						0	20	18	10	4	3	1
I (3rd theft)							0	23	14	7	3	1
F (late abortion)								0	28	24	20	1
H (unwanted sex)									0	18	15	2
K (cocaine dealer)										0	2	8
L (cocaine importer)											0	14
J (rape)												0

N = 64

NB: Shading indicates instances where adjacent scenarios were "flipped" in their rankings.

Each of the sixty-four subjects could have deviated from the group modal ranking as to each of the sixty-six pairwise judgments—the number of cells on the table—for a total number of 4224 possible deviations from modal pairwise ranking. Recall that if a subject ranks a single scenario different from the group's overall ranking it may produce many pairwise deviations. The subjects had a total of 561 deviations, meaning that the subjects reversed the overall group's pairwise ranking 13.3% of the time, as compared to the roughly 50% disagreement rate one would see for a random ranking of pairs.[223] This suggests more than three times the 4% disagreement rate among the same subjects when they ranked the first set of twenty-four scenarios in Study 1. A similar implication might be drawn from the fact that the standard deviations of the rank means in the second set of twelve scenarios are higher than those for the first set of twenty-four scenarios,[224] even though they should be lower by virtue of the smaller number of scenarios.

Again, this analysis of pairwise comparisons, rank-mode frequencies, and standard deviations of mean ranks is designed only to give the reader an overview of the data. The more reli-

223. Remember, if subjects were responding randomly, the subjects would agree in their pairwise comparison roughly 50% of the time.

224. Hand Run Standard Deviations Table (on file with the authors). The standard deviations of means in this context are particularly unreliable, however, because the subjects for Study 3 used their Study 1 rankings as their baseline.

able and sophisticated analysis is the use of Kendall's W coefficient of concordance, which measures agreement among sets of rank orderings. Recall that a W of 1.0 indicates perfect agreement, while 0.0 indicates no agreement.[225] For the twelve scenarios in Study 3, the Kendall's W is 0.55 ($p < 0.001$),[226] as compared to the Kendall's W of 0.95 for the same participants ranking the twenty-four Study 1 scenarios. Whether the disagreements shown here are influenced in part by demographic differences among subjects is an interesting question that we are planning to explore.[227]

3. Study 4 Results

Table 8 shows the mean rankings of the same twelve scenarios as ranked by the 246 subjects who did Internet-based rankings for Study 4. Remember that in this study, as opposed to paper-based Study 3, the subjects ranked the twelve scenarios among themselves, and not in relation to the twenty-four scenarios they had previously ranked for Study 2. Thus, the mean rankings in Table 8 can be compared to an ideal set of mean rankings from 1 to 12.

225. *See* ZAR, *supra* note 180, at 390.

226. N = 64; Chi-square = 389.39; df = 11.

227. Full analyses of this data would require sophisticated procedures beyond the scope of the present Article. Work is currently underway to answer the question of whether the demographic variables might predict differences in how scenarios are ranked. In particular, using a series of ordered probit regressions, we can determine for each offense whether the ranking of that offense varies as a function of the value of the demographic and other response variables collected during the survey. For example, because we collected data on gender, we can determine whether the gender of the subject correlates with the ranking given the "bestiality" scenario. By comparing the number of scenarios that are ranked statistically significantly different from one another as a function of questionnaire items, we can draw inferences about the extent to which demographics and political views influence evaluations of different offenses. We are planning to conduct further analyses on this dataset, Crime & Punishment Study, which is on file with the authors.

Table 8: Summary Data from Study 4

Scenario	Mean Rank*	Modal Rank	% Assigning "No Liability"
C (marijuana)	2.2	†	33
B (prostitution)	2.4	†	30
G (cocaine)	4.0	†	19
E (bestiality)	4.2	†	16
D (teen alcohol)	4.8	5	6
A (drunk crash)	6.2	6	0
I (3rd theft)	7.1	7	0
F (late abortion)	7.5	12	11
K (cocaine dealer)	7.9	9	6
H (unwanted sex)	8.7	11	1
L (cocaine importer)	8.9	10	6
J (rape)	11.1	12	0
N = 246			

* As noted previously, mean ranks in ordinal data obviously have limited usefulness but are provided simply to give a heuristic sense of the distribution for each scenario and to establish the relative relation among the scenarios for the layout of Table 9.

† These scenarios had a modal rank of "no punishment."

Using the mean group ranking in Table 8 to order the scenarios along each axis, Table 9 shows the frequencies with which subjects' judgments about the ranking of each scenario pair differed from that of the group's modal ranking of the pair.

Table 9: Deviations from Modal Ranks in Study 4

Scenario	C	B	G	E	D	A	I	F	K	H	L	J
C (marijuana)	0	99	25	84	51	39	22	26	5	19	5	6
B (prostitution)		0	52	74	56	39	24	29	11	17	10	5
G (cocaine)			0	122	107	78	48	51	9	35	6	6
E (bestiality)				0	97	74	54	59	50	32	44	10
D (teen alcohol)					0	88	63	68	34	31	22	9
A (drunk crash)						0	91	99	59	56	48	13
I (3rd theft)							0	110	82	67	58	14
F (late abortion)								0	122	107	115	62
K (cocaine dealer)									0	92	38	32
H (unwanted sex)										0	133	21
L (cocaine importer)											0	48
J (rape)												0

N = 246

NB: Shading indicates instances where adjacent scenarios were "flipped" in their rankings.

Each of the 246 subjects could have deviated from the group modal ranking as to each of the sixty-six pairwise judgments—the number of cells on the table—for a total number of

16,236 possible deviations from modal pairwise ranking. The subjects had a total of 3362 deviations, meaning that the subjects reversed the overall group's pairwise ranking 20.7% of the time, as compared to the roughly 50% disagreement rate one would see for a random ranking of pairs.[228] This result is noticeably higher than the 9.2% rate of disagreement among the same subjects when they ranked the first set of twenty-four scenarios, in Study 2.

Table 10 compares the standard deviation of the mean ranks of the twenty-four scenarios in Study 2 to the standard deviation of the mean ranks when the same 246 subjects ranked the twelve scenarios in Study 4. All other things being equal, for a scenario set of twelve scenarios in Study 4, the standard deviations for a scenario set of twenty-four scenarios in Study 3 should be roughly half of those for the smaller first set.

Table 10: Variation in Mean Ranks for Studies 2 and 4

24 Scenarios of Study 2	Standard Deviaion of Mean Rank	12 Scenarios of Study 4	Standard Deviation of Mean Rank
S1 (defending)	3.2	C (marijuana)	2.3
S2 (coercion)	5.4	B (prostitution)	2.4
S3 (umbrella)	0.9	G (cocaine)	2.6
S4 (hallucination)	3.3	E (bestiality)	3.4
S5 (pies)	2.4	D (teen alcohol)	2.5
S7 (short change)	3.0	A (drunk crash)	2.4
S6 (T-shirt)	2.5	I (3rd theft)	2.4
S8 (radio)	2.4	F (late abortion)	4.0
S9 (drill)	2.4	K (cocaine dealer)	2.6
S10 (microwave)	2.7	H (unwanted sex)	2.7
S11 (TV)	2.5	L (cocaine importer)	2.8
S12 (slap)	2.5	J (rape)	1.8
S13 (head-butt)	3.2		
S14 (stitches)	2.5		
S15 (necklace)	3.0		
S16 (robbery)	2.3		
S17 (clubbing)	2.4		
S18 (pit bulls)	2.1		
S19 (infant)	3.2		
S20 (stabbing)	1.6		
S21 (ambush)	1.5		
S22 (abduction)	1.5		
S23 (burning)	1.6		
S24 (ransom)	1.9		
Average SD	2.5		2.7

228. *See supra* note 179.

Rather than being half the standard deviations of the larger scenario set for Study 2, the standard deviations for the second set of twelve scenarios, are, in fact, higher. The first set has an average standard deviation of 2.5. The second set has an average standard deviation not of 1.25, but rather of 2.7.

Again, this analysis of pairwise comparisons, mean ranks, and standard deviation of mean ranks is designed only to give the reader a general overview of the data. The more reliable and sophisticated analysis is the use of Kendall's W coefficient of concordance, which measures agreement among sets of rank orderings. For the twelve scenarios in Study 4, the Kendall's W is 0.51 ($p < 0.001$).[229]

4. Discussion

As compared to the extremely high Kendall's W of 0.95 and 0.88 for the twenty-four scenarios in Studies 1 and 2, respectively, the Kendall's W for the twelve scenarios in these two studies, of 0.55 and 0.51, respectively, shows that both paper-based and Internet-based subjects have considerably less agreement on the second set of scenarios than on the first set. There is clearly more disagreement among subjects about the proper ranking of the Study 3 and 4 scenarios than about the proper ranking of the Study 1 and 2 scenarios. Why should this be so?

Each of the second set of offenses has its own story as to the reasons for its controversial nature.[230] But there may be

229. N = 246; Chi-square = 1391; df = 11. Whether the disagreements shown here are influenced in part by demographic differences among subjects is an interesting question that we plan to explore. *See supra* note 227.

230. For example, one arguing for enhanced drunk driving penalties may point to the number of serious automobile accidents caused by intoxicated drivers, while others may suggest that the existing punishments for impaired drivers are disproportionate to the sanctions for similarly reckless conduct, such as driving grossly over the speed limit in a residential area.

Prostitution can be characterized as either a victimless, consensual transaction or a serious problem that both disrupts the family and degrades the status of women. Likewise, those arguing for strong marijuana purchase sanctions may allude to the social harms caused by marijuana and its status as a gateway drug, while others will claim that marijuana use is a similarly victimless crime that causes no more social harm than the legal use of alcohol.

Respondents claiming that purchasing alcohol for a teenager is a minor offense will note the legal status of alcohol for older users and suggest that alcohol use by teenagers, when not combined with other conduct such as driving, causes no substantial harm. On the other hand, those supporting a more substantial sentence for such alcohol purchases will assert that teenagers are not yet old enough to use alcohol responsibly, and thus such other conduct occurs

systematic differences between the kind of scenarios of the first set and those of the second set that invite further exploration. Note, for example, that the second set of scenarios is different from the first set in that it concerns conduct outside the core of wrongdoing, which includes physical injury, taking without consent, and deception in exchanges. Such wrongs are so central to effective group cooperation that they may arise in any group in any culture. One might speculate that views about punishment of conduct outside the core of wrongdoing may not be the product of intuition at all, but rather the product of general social learning and reasoning. The more this is true, the more the situational context and the demographic characteristic of the subject may shape the punishment judgment. The results here suggest generally that the closer conduct is to the core of physical injury of persons or property, takings without consent, and deception in exchanges, the greater will be present-day agreement about its relative blameworthiness. The

with regularity once teenagers are provided with liquor.

Bestiality may be considered either private conduct and, in any case, a somewhat trivial offense, or a serious moral transgression and animal cruelty issue. If one believes that human life does not begin until birth, a late-term abortion may be considered nothing other than a valid exercise of a woman's right to control what happens to her body, while one who believes that life begins at an earlier point may equate a late-term abortion to homicide. Those suggesting that significant penalties should be associated with a cocaine purchase will note the often-violent behavior that can be attributed to cocaine users and the association of cocaine sellers with organized crime, while others will assert that cocaine use, in many contexts, is also a victimless crime.

One may argue that date rape is less serious than forceful rape and may claim that much of the punishment resulting from forcible rape can be attributed to the violence and threat associated with the act and analogize to the difference between theft and robbery. Others may suggest that the violation of a woman's dignity, regardless of whether violence is involved, is the same. Arguments in favor of an enhanced sentence after a third felony offense, such as jewelry grab, include the need to incapacitate an offender who seemingly cannot stop committing crimes and a claim that after a first and second conviction, one "should have learned," while claims against an enhanced sentence center on arguments for proportionality and that one's "punishment should fit the crime."

The arguments for and against stringent sanctions for large scale cocaine selling, and very large scale cocaine importation and distribution, are similar. Those in favor of harsher sentences will allude to the association between cocaine distribution and violent, organized crime and assert that even if cocaine use is a victimless crime that only harms the user, those distributing controlled substances are creating the victims. Those in favor of lighter sanctions for such offenders may take note of the dire economic conditions from which most cocaine distributors come and claim that the harms associated with cocaine use are exaggerated.

more conduct is tied to these fundamental wrongdoings only by analogy, the greater will be the disagreement over its seriousness, to an extent that reflects the disagreement over the strength of the analogy.[231]

CONCLUSION

As the previous sections have demonstrated, contrary to common wisdom, available evidence suggests that human intuitions of justice about core wrongdoing—both the sense that serious wrongdoing should be punished and the sense of the relative seriousness of wrongdoing—are deep, predictable, and widely shared. While there are disagreements about the relative blameworthiness of wrongdoing outside the core, the core wrongs themselves—physical aggression, takings without consent, and deception in exchanges—are the subject of nuanced and specific intuitions that cut across demographics.

The existence of such shared intuitions of justice may have important implications for criminal justice debates, including abolition of punishment, distribution of punishment according to principles that conflict with the community's shared intuitions of justice, and programs to change people's views on what constitutes serious wrongdoing. While a full account of these implications is another project,[232] consider the possibilities.

For example, it may be unrealistic to expect the population to all "rise above" its desire to punish wrongdoers, or to expect the government to "reeducate" people away from their interest in punishing wrongdoers, as is urged by some reformers. It seems unlikely that social engineering can change the shared intuition that serious wrongdoing should be punished, at least not through methods short of the kind of coercive indoctrination that liberal democracies would find unacceptable. If the intuition to punish serious wrongdoing were subject to easy manipulation, one would expect to find differences in this intuition among different contexts and demographics. Given the near universality of this intuition across societies and demographics, it is logical to assume that this intuition is insulated from the influence of the society and the situation.

231. We explore these issues further in Robinson, Kurzban & Jones, *supra* note 191.

232. *See* Paul H. Robinson & John M. Darley, *Intuitions of Justice: Implications for Criminal Law and Justice Policy*, 81 S. CAL. L. REV. (forthcoming 2008).

For another example, a criminal justice system that regularly fails to do justice or that regularly does injustice, as judged by the community's shared intuitions of justice, will inevitably be widely seen as failing in a mission that the community thinks important, or even foundational, unless the system's unjust operation can be hidden. Hiding injustice would be hard to do without breaching notions of press freedom and government transparency to which liberal democracies aspire.

For a final example, it seems likely that any realistic criminal justice system or program for its reform must take into account the community's shared intuitions of justice. This does not mean that law can never deviate from those intuitions or try to change them, but it must be realistic about what it can and cannot change and about the costs—financial and social—that such changes may require. The greatest success in shaping the perceived wrongfulness of particular conduct may not be to fight people's intuitions of justice but rather to try to harness them, by providing information or arguments that strengthen (or weaken) the analogy between the target conduct and the core wrongdoing on which people have strong intuitions.

APPENDIX A: TEXT OF STUDY 1 & 2 SCENARIOS; INSTRUCTIONS

The scenarios are listed in the order predicted, with first four scenarios having no liability.

SCENARIO 1: DEFENDING ATTACK

John is knocked down from behind by a man with a knife who moves to stab him. As the man lunges for him, John stabs him with a piece of glass he finds on the ground, which is the only thing he can do to save himself from being killed. The man later dies of his injuries.

SCENARIO 2: COERCIVE THREAT TO CHILD

A man grabs John's child and puts a sharp knife to her throat. He tells John that he will kill the child if John does not steal an expensive digital camera from a nearby shop or he attempts to contact police. Because the man can see everything he does, John does as he is told in order to save his child.

SCENARIO 3: UMBRELLA MISTAKE

John takes another person's umbrella assuming it to be his own because it is has the same unusual color pattern as his own, a fact that the police confirm.

SCENARIO 4: HALLUCINATION

Another person slips a drug into John's food, which causes him to hallucinate that he is being attacked by a wolf. When John strikes out in defense, he does not realize that he is in fact striking a person, a fact confirmed by all of the psychiatrists appointed by the state, who confirm that John had no ability to prevent the hallucination.

SCENARIO 5: WHOLE PIES FROM BUFFET

The owner has posted rules at his all-you-can-eat buffet that expressly prohibit taking food away; patrons can only take what they eat at the buffet. The owner has set the price of the buffet accordingly. John purchases dinner at the buffet, but when he leaves he takes with him two whole pies to give to a friend.

SCENARIO 6: LOGO T-SHIRT FROM STORE

John notices in a small family-owned music store a T-shirt with the logo of his favorite band. While the store clerk is preoccupied with inventory, John places the $15 T-shirt in his coat and walks out, with no intention of paying for it.

SCENARIO 7: SHORT CHANGE CHEAT

John is a cab driver who picks up a high school student. Because the customer seems confused about the money transaction, John decides he can trick her and gives her $20 less change than he knows she is owed.

SCENARIO 8: CLOCK RADIO FROM CAR

As he is walking to a party in a friend's neighborhood, John sees a clock radio on the backseat of a car parked on the street. Later that night, on his return from the party, he checks the car and finds it unlocked, so he takes the clock radio from the back seat.

SCENARIO 9: ELECTRIC DRILL FROM GARAGE

John does not have all the tools he needs for his workshop but knows of a family two streets over who sometimes leave unlocked the door to the detached garage next to their house. When he next sees his chance, he enters the detached garage through the unlocked door and takes a medium-size electric drill, intending to keep it forever.

SCENARIO 10: MICROWAVE FROM HOUSE

While a family is on vacation, John jimmies the back door to their house and steps into their kitchen. On the counter, he sees their microwave, which he carries away.

SCENARIO 11: SMASHING TV

While a family is away for the day, John breaks in through a bedroom widow and rummages through the house looking for valuables. He can only find an 18-inch television, which angers him. When he gets it outside, he realizes that it is an older model than he wants, so he smashes it onto the driveway, breaking it into pieces.

SCENARIO 12: SLAP & BRUISING AT RECORD STORE

A record store patron is wearing a cap that mocks John's favorite band. John follows him from the store, confronts him, then slaps him in the face hard, causing him to stumble. The man's face develops a harsh black and yellow bruise that does not go away for some time.

SCENARIO 13: HEAD-BUTT AT STADIUM

While attending a football game, John becomes angry as he overhears an opposing fan's disparaging remarks about John's team. At the end of the game, John sticks his face in the man's face and head-butts him, causing a black eye and a gash that requires two stitches to close.

SCENARIO 14: STITCHES AFTER SOCCER GAME

Angry after overhearing another parent's remarks during a soccer match in which John's son is playing, John approaches the man after the game, grabs his coffee mug, knocks him down, then kicks him several times while he is on the ground, knocking him out for several minutes and causing cuts that require five stitches.

SCENARIO 15: NECKLACE SNATCH AT MALL

As a woman searches her purse for car keys in a mall parking lot, John runs up and grabs her gold necklace but it does not break. He yanks the woman to the ground by her necklace, where she gashes her head, requiring stitches. John runs off without the necklace.

SCENARIO 16: ATTEMPTED ROBBERY AT GAS STATION

John demands money from a man buying gas at a gas station. When the man refuses, John punches the man several times in the face, breaking his jaw and causing several cuts that each require stitches. He then runs off without getting any money.

SCENARIO 17: CLUBBING DURING ROBBERY

To force a man to give up his wallet during a robbery attempt, John beats the man with a club until he relinquishes his wallet, which contains $350. The man must be hospitalized for two days.

SCENARIO 18: MAULING BY PIT BULLS

Two vicious pit bulls that John keeps for illegal dog fighting have just learned to escape and have attacked a person who came to John's house. The police tell John he must destroy the dogs, which he agrees to do but does not intend to do. The next day, the dogs escape again and maul to death a man delivering a package.

SCENARIO 19: INFANT DEATH IN CAR

John is driving to see a man about buying an illegal gun but must baby-sit his friend's toddler son. It occurs to him that it is too hot to safely leave the toddler in the car but he decides to leave him anyway and to return soon. He gets talking with the seller, however, and forgets about the toddler, who passes out and dies.

SCENARIO 20: STABBING

John is offended by a woman's mocking remark and decides to hurt her badly. At work the next day, when no one else is around, he picks up a letter opener from his desk and stabs her. She later dies from the wound.

SCENARIO 21: AMBUSH SHOOTING

John knows the address of a woman who has highly offended him. As he had planned the day before, he waits there for the woman to return from work and, when she appears, John shoots her to death.

SCENARIO 22: ABDUCTION SHOOTING

A woman at work reveals John's misdeeds to his employer, thereby getting him fired. John devises a plan to get even with her. The next week he forces the woman into his car at knife point and drives her to a secluded area where he shoots her to death.

SCENARIO 23: BURNING MOTHER FOR INHERITANCE

John works out a plan to kill his 60-year-old invalid mother for the inheritance. He drags to her bed, puts her in, and lights her oxygen mask with a cigarette, hoping to make it look like an accident. The elderly woman screams as her

clothes catch fire and she burns to death. John just watches her burn.

SCENARIO 24: RANSOM, RAPE, TORTURE & STRANGLING

John kidnaps an 8 year-old girl for ransom, rapes her, then records the child's screams as he burns her with a cigarette lighter, sending the recording to her parents to induce them to pay his ransom demand. Even though they pay as directed, John strangles the child to death to avoid leaving a witness.

STUDY 1 & 3 ADMINISTRATION PROCEDURES

The instructions to participants read:

> Imagine you are given complete discretion to punish (or not punish) John for his actions in each of the following scenarios.
>
> On the table in front of you, carefully order the scenarios from the least amount of punishment that you think John deserves, to the most amount of punishment you think he deserves. You may order the cards either horizontally or vertically on the table, whichever you prefer. Set aside all those scenarios in which you would not punish John. Rank order all remaining scenarios; do not allow any ties.
>
> Each scenario is independent of the others. Do not use details from one scenario in determining punishment for any other scenario. Base your answers solely on what you believe to be just, regardless of any laws. The only thing you should focus on is how much punishment John deserves. Take as much time as you need to rank order the cards given to you. Remember to carefully read each scenario. After you complete the sorting, review your rankings to be sure they are as you want them to be. When you are finished with this task, notify the administrator and await further instructions. Leave the cards spread on the table.
>
> SPECIAL NOTE: Anyone who would like to do so may do the exercise again in 30 days and, if they produce the same ranking as they do today, they will be paid $25.00. No one is obliged to do the exercise again. The administrator will provide a phone number to call for those who do want to do the exercise again.

After participants had completed the task, they were given the following instructions as "Part 2" of the study:

> Review the ranking of the scenarios that you made in Part 1. This time, consider each scenario in relation to the scenario above it in terms of punishment deserved, and the scenario below it in terms of punishment deserved to see if you still agree with the rank order of each scenario in relation to the one ranked above and below. Make any adjustments to order that you feel are necessary to reflect the relative amount of punishment John deserves. Leave the cards spread on the table.

The administrator instructions were as follows:

Give the participant the brief description of the survey before they agree to take the survey. Give the test only when and where the participant will have no interruptions or distractions and where there is a clear table large enough for 20 or more piles of cards.

(1) Give the participant the Part 1 Instruction and the Part 1 Scenario Cards. Ask them to tell you when they completed Part 1. Direct them to read and carefully follow the Part 1 Instructions and to take as much time as they need. Leave the person alone to sort the cards. You may check briefly to be sure they are doing the task as intended. They may arrange the cards either horizontally or vertically, whichever they prefer.

(2) When the person is finished with Part 1, quickly look over their Part 1 rankings. (This helps convince them that Part 2 really is a new test that deserves their careful attention.) Then give them the instructions for Part 2 and leave them alone.

(3) When the person is finished with Part 2, leave their cards as they placed them and give them the Part 3 Instructions and the Part 3 Scenario Cards. Instruct them to read and carefully follow the Part 3 Instructions and to take as much time as they need. Leave the person alone to sort and place the cards. You may check briefly to be sure they are doing the task as intended.

(4) When the person has finished Part 3, ask them to complete the Demographics Questionnaire. Ask them to place the form in the "Demographics Forms" envelope when they are finished. (This helps avoid any awkwardness in having them hand the form with its personal information to you.) When they are finished, thank them for their help.

(5) Have them fill out the Charities form, or pay them $10 and sign a cash receipt.

(6) Record their responses on the Survey Response Form: Record their card rank ordering, including a horizontal line that distinguishes the no-liability cases. The rank of each Part 3 scenario should marked by a line drawn to even with or between two Part 1 scenarios.

Record on the bottom of the demographic form whether the subject was from Urban, Suburban, or Rural. Record on the demographic form whether the person had been used in field testing

Separate the Part 1 and Part 3 cards. Shuffle each deck to get ready for the next participant.

APPENDIX B: REASONS FOR PREDICTED RANKINGS OF STUDY 1 & 2 SCENARIOS

SCENARIOS 1–4: NO LIABILITY

Each of these scenarios presents a case of clear and compelling exculpation even though what John has done is normally prohibited. In scenario 1, the killing is necessary to defend against an unprovoked attempt to kill John, typically giving a complete justification. In scenario 2, the apparently inescapable threat to kill John's child justifies, or at least excuses, his coerced theft. In scenario 3, the honest and non-culpable mistake undercuts the culpable state of mind normally a prerequisite for criminal liability. In scenario 4, the hallucination, which John did not culpably cause, means that John lacks the kind of moral functioning that typically is a prerequisite for criminal liability, and thus excuses him.

SCENARIOS 5–11: THEFT AND INTRUSION

Scenario 5, in which John takes two pies from an all-you-can-eat buffet, was designed to be a borderline case. His conduct is technically in violation of the rules of the buffet but people could conclude that it is a violation insufficiently serious to merit the condemnation of criminal liability. The remaining scenarios in the series, 6 through 11, move along several continuums that appear to be important to people's judgments about the extent of such wrongdoing. The value of the property taken increases steadily from a T-shirt through an electric drill to a television. The extent of intrusion used in the taking also increases steadily from a public place through a person's unlocked car to rummaging through a person's house.

SCENARIOS 12–14: INJURY

In a fashion parallel to the technique used above, these injury scenarios 12 through 14 move along the continuums that people apparently use to judge the extent of wrongdoing in causing injury. The extent of the injury increases across the cases from a slap with brief bruising through a head-

butt to severe kicks requiring stitches. The overall sense of hostility and aggression also increases over the series.

SCENARIOS 15–17: ROBBERY

These robbery cases, 15 through 17, are in essence a continuation of the injury series above with the added wrongdoing of a forceful taking. Again, the extent of the injury increases from a fall causing a gash through punches causing a broken jaw to a clubbing requiring hospitalization. (The value of the goods sought to be taken also increases through the series.)

SCENARIOS 18 & 19: UNINTENTIONAL KILLINGS

Cases 18 and 19 present instances of greater harms than the injury and robbery series above but are distinguishable from the cases in the next series because here John has no intention to cause the death. Level of culpable state of mind toward a result is seen as a powerful determinant of blameworthiness. The greater punishment is predicted for the infant death over the adult because of the more vulnerable victim, also a powerful determinant in many contexts.

SCENARIOS 20–24: INTENTIONAL KILLINGS

These cases of intentional killings, 20 through 24, move along a continuum of greater brutality and greater planning and calculation in the execution of the offense, together with more vulnerable victims at the extreme.

APPENDIX C: TEXT OF STUDY 3 & 4 SCENARIOS

SCENARIO A: DRUNK DRIVING CRASH

John stops by Earl's Tavern on his way home from work, and drinks two of their infamous long island iced teas. Driving home he crashes his pickup into a telephone pole in his suburban subdivision, suffering only minor injuries. His blood alcohol content is twice the legal limit.

SCENARIO B: PROSTITUTION

The front desk clerk suspects Jane is a prostitute when she checks into the motel. After witnessing several different men enter and leave the motel room, he calls police, who arrive and witness Jane having sex for money. How much punishment does Jane deserve?

SCENARIO C: MARIJUANA PURCHASE FOR USE

John approaches a drug dealer and purchases enough marijuana to smoke six "bowls" for a party one Saturday night.

SCENARIO D: ALCOHOL FOR TEEN PARTY

John's 15 year-old cousin is throwing a party for all of his high school friends. Because his cousin is too young to legally buy alcohol himself, John buys him 2 kegs and a half-gallon of vodka, in violation of state law.

SCENARIO E: BESTIALITY

A local farmer wakes up in the middle of the night to strange noises coming from the barn. As he gets up to investigate, he discovers John having vigorous intercourse with one of the farmer's sheep.

SCENARIO F: LATE-TERM ABORTION

John is a general practitioner who is asked by one of his long-time patients to perform a late term abortion in the 7th month of pregnancy because she no longer wants the child. Although he knows such a late abortion is illegal, he performs the procedure anyway.

SCENARIO G: COCAINE PURCHASE FOR USE

John is picked up by the police after buying half a gram of powder cocaine on the corner of the street across the city from his home, enough for 6 "lines."

SCENARIO H: UNCONSENTED-TO INTERCOURSE

During an intimate petting session with his girlfriend, John gets sexually excited and asks for sex. His girlfriend says no, and protests as he climbs on top of her. She lays stiffly, without moving, and she tells him to stop, but he does not.

SCENARIO I: JEWELRY GRAB THIRD CONVICTION

John asks to examine a diamond engagement ring at a store, then quickly runs off with it, but is caught. John has two previous convictions for which he served jail time, one for stealing a car and the other for starting a fist fight with another man.

SCENARIO J: ASSAULT & FORCIBLE INTERCOURSE

John hides in the bushes near a burned out street lamp in a mall parking lot. When a mall employee returns to her car after work John drags her back into the bushes and rapes her.

SCENARIO K: DRUG DEALER

John earns a lot of money as one of the only drug dealers in town, selling cocaine to anyone who will buy. Police raid his expensive apartment, and find 500 grams of powder cocaine under his basement, enough for around 6,000 "lines."

SCENARIO L: COCAINE IMPORTER/DISTRIBUTOR

John is a big time cocaine importer and distributor who lives in a beautiful mansion and directs the work of a dozen dealers. The police raid his home and find five kilograms of powder cocaine, enough for 60,000 "lines" of cocaine.

APPENDIX D: EXAMPLES OF STATUTORY MAXIMUMS FOR STUDY 3 & 4 SCENARIOS

DRUNK DRIVING CRASH

625 ILL. COMP. STAT. ANN. 5/11-501(b-2) (West 2006) (codifying the offense as an Illinois Class A Misdemeanor); *id.* at 5/11-501(b-3) (providing a mandatory minimum of 5 days imprisonment or 240 hours of community service for any second conviction).

GA. CODE ANN. § 40-6-391(a)(5) (2003) (providing offense); *id.* § 40-6-391(c)(1)(A)–(C) (providing punishment comprising a fine of $300–$1000, jail time of 10 days to 12 months, and community service of 40 hours).

PROSTITUTION

720 ILL. COMP. STAT. ANN. 5/11-14(b) (West 2006) (codifying the offense as an Illinois Class A Misdemeanor), *declared unconstitutional by* People v. Lindsey, 753 N.E.2d 1270 (Ill. App. Ct. 2001).

GA. CODE ANN. § 16-6-9 (2003) (providing offense); *id.* § 16-6-13(a) (codifying the offense as a misdemeanor).

MARIJUANA PURCHASE FOR USE

U.S. SENTENCING GUIDELINES MANUAL § 2D1.1(c)(17) (2006) (assigning Level 6 for less than 250 grams).

720 ILL. COMP. STAT. ANN. 550/4(a) (West 2006) (classifying the offense as an Illinois Class C Misdemeanor for not more than 2.5 grams).

GA. CODE ANN. § 16-13-30(j)(1) (2003) (providing offense); *id.* § 16-13-30(j)(2) (prescribing felony sentence of 1–10 years).

ALCOHOL FOR TEEN PARTY

235 ILL. COMP. STAT. ANN. 5/16-16(a) (West 2006) (codifying the offense as an Illinois Class A Misdemeanor), *declared unconstitutional by* People v. Law, 782 N.E.2d 247 (Ill. 2002); *id.* (providing punishment as a fine not less than $500).

GA. CODE ANN. § 3-3-23(a)(1)–(4) (2003) (providing offense); *id.* § 3-3-23.1(b)(1) (codifying the offense as a misdemeanor).

BESTIALITY

720 ILL. COMP. STAT. ANN. 5/12-35(e) (West 2006) (codifying the offense as an Illinois Class 4 Felony).

GA. CODE ANN. § 16-6-6(b) (2003) (providing punishment as 1–5 years imprisonment).

LATE-TERM ABORTION

720 ILL. COMP. STAT. ANN. 513/10 (West 2006) (codifying the offense as an Illinois Class 4 Felony), *declared unconstitutional by* Hope Clinic v. Ryan, 995 F. Supp. 847 (N.D. Ill. 1998).

GA. CODE ANN. § 16-12-140 (2003) (providing punishment of 1–10 years imprisonment).

COCAINE PURCHASE FOR USE

U.S. SENTENCING GUIDELINES MANUAL § 2D1.1(c)(14) (2006) (assigning Level 12 for less than 25 grams).

720 ILL. COMP. STAT. ANN. 570/402(a) (West 2006) (codifying the offense as an Illinois Class 1 Felony); *id.* at 570/402(a)(2)(A) (providing punishment as 4–15 years imprisonment with respect to 15–100 grams).

GA. CODE ANN. § 16-13-30(c) (2003) (providing a first offense sentence of 2–15 years imprisonment; a second offense sentence of 5–30 years imprisonment).

UNCONSENTED-TO INTERCOURSE

720 ILL. COMP. STAT. ANN. 5/12-13(b)(1) (West 2006) (codifying the offense as an Illinois Class 1 Felony Criminal Sexual Assault). It is unclear, however, whether the situation as depicted in the scenario would be a crime under the Illinois Code, as John did not "commit[] an act of sexual penetration *by the use of force or threat of force*." *Id.* at 5/12-13(a)(1) (emphasis added). There was no consent, but because John's girlfriend "lay[] stiffly, without moving," it is unclear that liability would be assigned. *See supra* Appendix C (Scenario H).

GA. CODE ANN. § 16-6-1(a) (2003) (providing offense when "carnal knowledge" is acquired "forcibly and against her will"); *id.* § 16-6-1(b) (providing offender "shall be punished by death, by imprisonment for life without parole, by imprisonment for life, or by . . . a term of imprisonment for not less than 25 years and not exceeding life imprisonment").

In Georgia, the offense of rape requires more than non-consensual sex; it requires the element of force. Perry v. State, 588 S.E.2d 838, 840 (Ga. Ct. App. 2003). If the State establishes beyond a reasonable doubt that an alleged rape victim's lack of outward or physical resistance resulted from her apprehension of bodily harm, violence, or other dangerous consequences to herself or another, that lack of resistance will not constitute "freely given consent." Clark v. State, 404 S.E.2d 787, 788 (Ga. 1991). "Force, as an element of rape, need not be proven by evidence of physical violence; evidence of a victim's lack of resistance induced by fear authorizes a finding of force." *In re* J.W.L., 531 S.E.2d 169, 170 (Ga. Ct. App. 2000). It is unclear whether the force element would be met here because it is unclear that the victim's lack of resistance was induced by fear. *See supra* Appendix C (Scenario H).

JEWELRY GRAB THIRD CONVICTION

720 ILL. COMP. STAT. ANN. 5/16-1(b)(2) (West 2006) (providing that theft of property from the person not exceeding $300 in value, or theft of property exceeding $300 and not exceeding $10,000 in value, is an Illinois Class 3 Felony).

GA. CODE ANN. § 16-8-40(a)(3) (2003) (providing a robbery offense for "sudden snatching"); *id.* § 16-8-40(b) (providing a sentence of 1–20 years imprisonment); *id.* § 17-10-7(a) (providing that for a second felony conviction the sentence is presumptively the maximum time in the underlying offense).

ASSAULT & FORCIBLE INTERCOURSE

720 ILL. COMP. STAT. ANN. 5/12-13(b)(1) (West 2006) (codifying the offense as an Illinois Class 1 Felony Criminal Sexual Assault).

GA. CODE ANN. § 16-6-1(a) (2003) (providing an offense when "carnal knowledge" is acquired "forcibly and against her will"); *id.* § 16-6-1(b) (providing that the offender "shall be punished by death, by imprisonment for life without parole, by imprisonment for life, or by . . . imprisonment for not less than 25 years and not exceeding life imprisonment").

DRUG DEALER

720 ILL. COMP. STAT. 570/401(a) (West 2006) (codifying the offense as an Illinois Class X Felony); *id*. at 570/401(a)(2)(C) (providing punishment as 12–50 years imprisonment with respect to 400–900 grams of a substance containing cocaine).

GA. CODE ANN. § 16-13-31(a)(1) (2003) (providing for a minimum sentence of 25 years imprisonment and a fine of $1 million for "knowingly sell[ing]" or being "knowingly in possession of" greater than 400 grams of cocaine).

COCAINE IMPORTER/DISTRIBUTOR

U.S. SENTENCING GUIDELINES MANUAL § 2D1.1(c)(4) (2006) (assigning Level 32 for 5–15 kilograms).

720 ILL. COMP. STAT. ANN. 570/401(a)(2)(D) (West 2006) (providing punishment as 15–60 years imprisonment for 900 grams or more).

GA. CODE ANN. § 16-13-31(a)(1) (2003) (providing for a minimum sentence of 25 years and a fine of $1 million for "knowingly sell[ing]" or being "knowingly in possession of" greater than 400 grams of cocaine).