View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by St. John's University School of Law

St. John's Law Review

Volume 80
Number 4 *Volume 80, Fall 2006, Number 4*

Article 3

February 2012

The Stakes Matter: Empirical Evidence of Hypothetical Bias in Case Evaluation and the Curative Power of Economic Incentives

Daniel R. Cahoy

Min Ding

Follow this and additional works at: https://scholarship.law.stjohns.edu/lawreview

Recommended Citation

Cahoy, Daniel R. and Ding, Min (2006) "The Stakes Matter: Empirical Evidence of Hypothetical Bias in Case Evaluation and the Curative Power of Economic Incentives," *St. John's Law Review*: Vol. 80 : No. 4 , Article 3.
Available at: https://scholarship.law.stjohns.edu/lawreview/vol80/iss4/3

This Article is brought to you for free and open access by the Journals at St. John's Law Scholarship Repository. It has been accepted for inclusion in St. John's Law Review by an authorized editor of St. John's Law Scholarship Repository. For more information, please contact selbyc@stjohns.edu.

THE STAKES MATTER: EMPIRICAL EVIDENCE OF HYPOTHETICAL BIAS IN CASE EVALUATION AND THE CURATIVE POWER OF ECONOMIC INCENTIVES

DANIEL R. CAHOY† AND MIN DING††

INTRODUCTION

Jury research plays a critical role in the modern legal environment. In the private dispute context, trial consulting companies commonly run mock trial simulations in order to determine the effect of facts or issues particular to a client's case.[1] Additionally, a growing number of courts employ a technique known as a summary jury trial that makes use of a surrogate jury to provide information on the relative strength of each party's case in order to motivate settlement.[2] In the

---

† Assistant Professor of Business Law, Smeal College of Business, the Pennsylvania State University; J.D. 1998, Franklin Pierce Law Center; B.A. 1991, University of Iowa. The authors acknowledge financial support from the Smeal Research Grant Program, and helpful comments from Gary Bolton, Keith Crocker, Robert A. Prentice, Abdullah Yavas, as well as the participants of the Academy of Legal Studies in Business Annual Meeting (2004) and the Smeal Faculty Colloquia. The method described in this paper is currently the subject of a pending U.S. patent application. *See* U.S. Patent Pub. No. US 2006/0036464 A1 (filed Aug. 10, 2005) (Cahoy & Ding, inventors).

†† Assistant Professor of Marketing, Smeal College of Business, the Pennsylvania State University; Ph.D. 2001, The Wharton School of the University of Pennsylvania; Ph.D. 1996, Ohio State University; B.S. 1989, Fudan University.

[1] *See* Franklin Strier & Donna Shestowsky, *Profiling the Profilers: A Study of the Trial Consulting Profession, Its Impact on Trial Justice, and What, If Anything, To Do About It*, 1999 WIS. L. REV. 441, 456 (1999) ("Consultants report spending the greatest percentage of their consulting time on mock trial simulations, followed by case theory/presentation and focus groups, respectively."); Stephanie Leonard Yarbrough, *The Jury Consultant—Friend or Foe of Justice*, 54 SMU L. REV. 1885, 1893–94 (2001) (noting that "[m]ock trials are one of the most popular instruments used by jury consultants.").

[2] *See* Harry T. Edwards, *Alternate Dispute Resolution: Panacea or Anathema?*, 99 HARV. L. REV. 668, 673 & n.16 (1986) (stating that a "summary jury trial" was introduced by Judge Lambros and was modeled after the mini-trial).

academic context, experimental studies on jury behavior are undertaken to uncover the biases, preconceptions and emotional triggers that influence the behavior of juries.[3] Such studies are critically important to the judicial system for determining the best ways to mitigate (or at least anticipate) the effects of racism, sexism, and economic bias.[4]

However, the use of simulated or "mock" juries has serious limitations and disadvantages. Most significantly, researchers in the field acknowledge the possibility that the hypothetical nature of jury simulation studies leads to subject behavior that differs from that of real jurors.[5] This so-called "consequentiality" or "hypothetical bias" manifests as a barrier to eliciting reliable responses from participants in a laboratory setting.[6] It occurs because the real-world decision-making incentives that flow from the impact of the decision are lacking. Simply put, a participant may make choices other than what he or she would if the study conditions were real; the stakes can matter, and the failure to account for them can be very problematic. Interestingly, the potential for skewed results in jury studies is concern enough that even the U.S. Supreme Court has commented on the

---

[3] *See* Daniel R. Cahoy & Min Ding, *Using Experimental Economics To Peek into the "Black Box" of Jury Behavior: A Proposal for Jury Research Reform*, 14 S. CAL. INTERDISC. L.J. 31, 37, 44, 48 (2004) (discussing the reasons for conducting jury research and the problems associated with those studies).

[4] *See id.* at 37.

[5] *See* NORMAN J. FINKEL, COMMONSENSE JUSTICE 58–61 (1995) (discussing the advantages and disadvantages of the experimental method); ROBERT J. MACCOUN, GETTING INSIDE THE BLACK BOX: TOWARD A BETTER UNDERSTANDING OF CIVIL JURY BEHAVIOR 14 (1987); Brian H. Bornstein & Sean G. McCabe, *Jurors of the Absurd? The Role of Consequentiality in Jury Simulation Research*, 32 FLA. ST. U. L. REV. 443, 444, 448, 455 (2005) (stating from a theoretical perspective that "one limitation seems insurmountable, as it is the *sine qua non* of a simulation; namely, no matter how realistic a simulation is, it is still just a simulation," and describing academic criticism of the hypothetical nature of jury studies); Ronald Dillehay & Michael Nietzel, *Constructing a Science of Jury Behavior*, *in* REVIEW OF PERSONALITY AND SOCIAL PSYCHOLOGY 253–54 (L. Wheeler ed., 1980) (discussing the problems with common jury research methods and stating that "[w]e are skeptical that the prototypical jury analogue is capable of capturing the complex socialization processes which are produced by the jury experience"); Shari Seidman Diamond, *Illuminations and Shadows from Jury Simulations*, 21 LAW & HUM. BEHAV. 561, 566 (1997) (concluding after reviewing the results of more detailed modern jury studies that "we can take some comfort that our efforts to invest more resources in the jury simulation paradigm have been justified," and acknowledging that "even these more elaborate simulations cannot avoid some of the inevitable uncertainties that research can reduce, but not avoid").

[6] *See* Cahoy & Ding, *supra* note 3, at 48.

problem.[7] While the actual impact of hypothetical bias on the reliability of mock jury studies is open to question,[8] it is clear that research into the issue is necessary and relevant.

The phenomenon of hypothetical bias is well characterized in the experimental literature of many social science fields.[9] Not surprisingly, various methods of reducing its effects have been developed.[10] Generally, the standard model involves the use of a reward or compensation that is directly linked to a participant's response.[11] Unfortunately, existing economic models for such amelioration are not useful in the context of jury studies due to their unique incentive structure. Specifically, juries[12] are not

---

[7] *See, e.g.*, Lockhart v. McCree, 476 U.S. 162, 171 (1986) (criticizing jury studies cited to the court, Chief Justice Rehnquist noted that they "were based on the responses of individuals randomly selected from some segment of the population, but who were not actual jurors sworn under oath to apply the law to the facts of an actual case involving the fate of an actual capital defendant" and that "[w]e have serious doubts about the value of these studies in predicting the behavior of actual jurors").

[8] *See* Robert J. MacCoun, *Comparing Legal Factfinders: Real and Mock, Amateur and Professional*, 32 FLA. ST. U. L. REV. 511, 513–14 (2005) (conceding the possibility that mock juror behavior differs from real juror behavior, but arguing that there is little evidence that the differences are asymmetric rather than in magnitude); Michael J. Saks, *What Do Jury Experiments Tell Us About How Juries (Should) Make Decisions?*, 6 S. CAL. INTERDISC. L.J. 1, 7–8 (1997) (noting that mock jury studies elicit the most concern of all jury experiments, but arguing that such worries are "out of proportion to the problems" of such research given that effects from the methodologies employed are not necessarily relevant to the variable under study).

[9] *See* Richard C. Bishop & Thomas A. Heberlein, *Does Contingent Valuation Work?*, *in* RONALD CUMMINGS ET AL., VALUING ENVIRONMENTAL GOODS: A STATE OF THE ARTS ASSESSMENT OF CONTINGENT VALUATION METHOD 149, 151 (1986) (analyzing the effect hypothetical bias had on several hunters' choices whether to sell their permits); Min Ding et al., *Incentive Aligned Conjoint Analysis*, 42 J. MARKETING RES. 67, 67–68 (2005) (testing hypothetical bias through a field experiment in a Chinese restaurant); John A. List, *Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards*, 91 AM. ECON. REV. 1498, 1501–02 (2001) (analyzing the effect of hypothetical bias in a sportscard auction experiment); John A. List & Jason F. Shogren, *The Deadweight Loss of Christmas: Comment*, 88 AM. ECON. REV. 1350, 1354 (1998) (studying the effect hypothetical bias can have on Christmas gift purchases).

[10] *See* List, *supra* note 9, at 1498.

[11] Alvin E. Roth, *Introduction to Experimental Economics*, *in* THE HANDBOOK OF EXPERIMENTAL ECONOMICS 6 (John Kagel & Alvin Roth eds., 1995) (noting that in as early as the 1940s, criticism regarding hypothetical choices in economics experiments became widely acknowledged as a result of a paper by W. Allen Wallis and Milton Friedman, and since that time the use of real incentives have become a fundamental aspect of experimental economics).

[12] The term "jury" is used herein to refer to petit juries as opposed to grand

motivated by the potential for personal gain, but instead are called upon to make just decisions for others. A novel approach to capture this incentive is necessary.

This paper proposes a remedy for the problem of hypothetical bias in jury studies that is translatable to existing modalities. We begin in Part I by providing background from the relevant literature, exposing hypothetical bias in other, limited contexts. In Part II, the paper describes a mechanism specifically designed to align incentives in jury studies, articulating its economic basis and components. Finally, in Part III, the paper presents the results of two experiments that demonstrate the utility of the incentive structure and provides a roadmap for future work in this area.

I. DESCRIPTIONS OF HYPOTHETICAL BIAS IN THE SOCIAL SCIENCE LITERATURE AS A PARTIAL ROADMAP FOR A REMEDY

Whenever an individual is called upon to essentially predict what he or she would do in a given circumstance, there is a potential for hypothetical bias. In the context of jury simulation research, the phenomenon is certainly acknowledged, but infrequently studied.[13] According to a recent survey of the literature, only five studies have directly addressed the issue, and the results are now at least twenty years old.[14] Unfortunately, this limited body of research is ultimately

juries. A petit jury is a panel of persons from the community empanelled to determine factual issues in the context of a trial, a long-standing tradition in the United States. *See* Colgrove v. Battin, 413 U.S. 149, 176–77 (1973) (Marshall, J., dissenting) (noting that the use of petit juries has been described at least as early as 1164).

[13] *See* MacCoun, *supra* note 8, at 512 (observing that "[r]eaders outside the jury research community might be surprised to find so little attention given to . . . the consequentiality issue").

[14] *See* Bornstein & McCabe, *supra* note 5, at 452 & n.39. Noted researcher, Brian Bornstein, led this comprehensive assessment of the issue of consequentiality in jury studies in 2005, and it provides a helpful picture of the state of the research to date. The studies cited by Bornstein and McCabe are: Shari Seidman Diamond & Hans Zeisel, *A Courtroom Experiment on Juror Selection and Decision-Making*, 1 PERSONALITY & SOC. PSYCHOL. BULL. 276 (1974); David W. Wilson & Edward Donnerstein, *Guilty or Not Guilty? A Look at the "Simulated" Jury Paradigm*, 7 J. APPLIED SOC. PSYCHOL. 175 (1977); Norbert L. Kerr et al., *Role Playing and the Study of Jury Behavior*, 7 SOC. METHODS & RES. 337 (1979); David Suggs & John J. Berman, *Factors Affecting Testimony About Mitigating Circumstances and the Fixing of Punishment*, 3 LAW & HUM. BEHAV. 251 (1979); and Martin F. Kaplan & Sharon Krupa, *Severe Penalties Under the Control of Others Can Reduce Guilt Verdicts*, 10 LAW & PSYCHOL. REV. 1 (1986).

inconclusive.[15] Four of the five studies found a direct or interacting effect of role-playing and consequences on jury behavior,[16] one found no effect,[17] and all suffer from some methodological shortcomings that limit the objective reliability of the results.[18]

On the other hand, a wealth of empirical evidence on hypothetical bias exists in the experimental literature related to the value subjects place on certain options or contingencies. Diamond and Hausman[19] reviewed surveys that purported to gauge how much value people placed on public goods, such as cleaning up polluted rivers and lakes.[20] They concluded that the discrepancy between the valuation stated under hypothetical conditions and real-life responses was so profound that information gleaned from contingent valuation surveys was actually *worse* than no information at all.[21] Additionally, in a study by Bishop and Heberlein, it was found that the amount of money people were willing to spend for deer hunting permits was overstated when compared to a situation in which they would have to actually spend cash.[22] List showed that sports card dealers significantly overstate their bids for a sports card in a hypothetical situation.[23] List and Shogren found the selling price for a gift is significantly higher in an actual situation as compared to a hypothetical.[24] Moreover, Ding, et al., recently found that subjects were more price-sensitive and had a different

---

[15] *See* Bornstein & McCabe, *supra* note 5, at 457 ("The results of the above studies provide little general consensus about the effect of role-playing and consequences on jury behavior.").

[16] *See* Diamond & Zeisel, *supra* note 14, at 276–77; Kaplan & Krupa, *supra* note 14, at 8–13; Suggs & Berman, *supra* note 14, at 256; Wilson & Donnerstein, *supra* note 14, at 185.

[17] *See* Kerr et al., *supra* note 14, at 348.

[18] *See* Bornstein & McCabe, *supra* note 5, at 457–60 (identifying such failings as the lack of evidence that hypothetical bias had been manipulated successfully and the failure to employ jury deliberations).

[19] Peter A. Diamond & Jerry A. Hausman, *Contingent Valuation: Is Some Number Better than No Number?*, 8 J. ECON. PERSP. 45 (1994).

[20] *Id.* at 45–46.

[21] *See id.* at 58–60 (discussing the "some number is better than no number" fallacy).

[22] *See* Bishop & Heberlein, *supra* note 9, at 157–58.

[23] *See* List, *supra* note 9, at 1501–02.

[24] List & Shogren, *supra* note 9, at 1354.

preference structure for Chinese dinner specials when their decisions resulted in actually eating the food.[25]

Various behaviors have been posited to explain hypothetical bias. For example, participants in hypothetical settings may be shifting their behavior to an overly socially-desirable presentation of themselves: "[w]hen incentives are low subjects say they would be more risk-preferring and generous than they actually are when incentives are increased."[26] Additionally, there may be less heterogeneity among the respondents in hypothetical situations (e.g., they may be more likely to conform to the social norm or to answers expected by their peers).[27] In general, answers under hypothetical conditions are inconsistent, erratic, and, in many cases, untrustworthy.[28] Finally, participants may discount things that would be important in real-world contexts (e.g., budget constraints).[29]

To correct the hypothetical bias, non-legal experimenters normally incorporate incentives that parallel their real-world counterparts, and motivate the respondents by linking their decisions to a reward amount.[30] However, jury research has often utilized a different approach to address the lack of realism. The standard practice, favored by research psychologists and jury consultants, is to make the experimental conditions as real as possible (e.g., by using live actors, formal settings, etc.) while allowing the consequences of the participants' decision to remain hypothetical.[31] The idea is that more realistic surroundings will

---

[25] Ding et al., *supra* note 9, at 70.

[26] Colin F. Camerer & Robin M. Hogarth, *The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework*, 19 J. RISK AND UNCERTAINTY 7, 8 (1999) (analyzing a meta study of seventy-four research papers which explored the impact of various financial incentives on the behavior of experimental subjects).

[27] *See id.* at 8–9.

[28] *See, e.g., id.* at 8; Vernon L. Smith & James M. Walker, *Monetary Rewards and Decision Cost in Experimental Economics*, 31 ECON. INQUIRY 245, 251–52 (1993) (reviewing experimental literature involving varying incentive structures in hypothetical situations).

[29] Vernon L. Smith, *Experimental Economics: Induced Value Theory*, 66 AM. ECON. REV. 274, 277–78 (1976) (describing an experiment, in which, subjects imagining themselves making a five cent commission on each contract was "not real enough to induce many contracts").

[30] *See supra* note 11 and accompanying text.

[31] *See* FINKEL, *supra* note 5, at 59–60; Cahoy & Ding, *supra* note 3, at 48; Diamond, *supra* note 5, at 561–62 (noting the use of more realistic study techniques have been substantially adopted in most modern jury studies).

elicit more realistic responses by encouraging participants to forget (or at least downplay) the fact that the consequences are not real.[32] Unfortunately, there is no guarantee that subjects will behave more closely to real life simply by virtue of increased realism.[33] Furthermore, this may turn out to be inconsequential.[34] A 1999 meta-analysis of twenty years of jury simulation research concluded that such measures appear to have little systematic effect.[35] More to the point, they do not directly address the hypothetical bias issue.[36] In fact, given the lack of any incentive to compel such behavior it is reasonable to assume that at least some participants will continue to deviate from what their real life response would be;[37] in even the most realistic studies, a very high level of hypothetical bias may still be present.

Some researchers have adopted an incentive-aligned methodology for addressing hypothetical bias that rewards individuals based on the outcome of their decision.[38] Guarnaschelli, McKelvey, and Palfrey, for example, conducted an experimental study of the rules underlying jury voting behavior using opaque jars of colored balls representing the state of the world: blue (which was analogized to innocent) and red

---

[32] SAUL M. KASSIN & LAWRENCE S. WRIGHTSMAN, THE AMERICAN JURY ON TRIAL: PSYCHOLOGICAL PERSPECTIVES 18 (1988) ("[W]e can say, as a general rule, that the more closely our research conditions approximate the real event, the better off we are trying to generalize from the former to the latter."); Cahoy & Ding, *supra* note 3, at 48–49.

[33] *See* Bornstein & McCabe, *supra* note 5, at 448 ("[E]xperimental methodologies used by researchers are becoming increasingly sophisticated and legally realistic. Nonetheless, an experiment is ultimately still an experiment, raising the issue of whether any simulation can meaningfully speak to real-world legal questions.").

[34] *See* MacCoun, *supra* note 8, at 512 ("It is difficult to avoid the conclusion that efforts to maximize realism on these dimensions have more to do with research marketing than scientific validity . . . .").

[35] Brian H. Bornstein, *The Ecological Validity of Jury Simulations: Is the Jury Still Out?*, 23 L. & HUM. BEHAV. 75, 76–84, 88 (1999) ("Two decades of additional research support Bray & Kerr's conclusion, regarding the ecological validity of jury simulations, that 'the pattern of results does not warrant the negative reactions of some evaluators.' ").

[36] Bornstein himself concedes that, while his 1999 study asserted that the ecological validity of simulation studies supported their continued use, it "turned a blind eye" to the consequentiality issue (as does most other jury simulation research). Bornstein & McCabe, *supra* note 5, at 450 & n.32.

[37] *See* Smith, *supra* note 29, at 274–79 (1976).

[38] *See* Serena Guarnaschelli et al., *An Experimental Study of Jury Decision Rules*, 94 AM. POL. SCI. REV. 407, 408 (2000).

(analogized to guilty).[39] One would obtain a "signal" as to the overall composition of the jar by individually viewing the color of one ball from the jar and voting accordingly.[40] The incentive to vote correctly was provided by the fact that, for each round of voting, all jurors were paid fifty cents for a correct decision on the composition of the jar and five cents for an incorrect one.[41]

It is clear, however, that straight application of personal incentives in jury simulation study may produce misleading results. This is because the missing incentive in jury simulation is not personal utility (as in the contexts commonly studied in the hypothetical bias literature), but rather social utility. Real jurors do not gain money by delivering a well-thought-out verdict, nor do they lose money even if they deliver a random verdict.[42] Real jurors derive significant utility knowing that, if they make the right decision, both the plaintiff and defendant will receive what they truly deserve.[43]

This distinction is no small matter. The social utility literature suggests such "non-self" utility can be critical in shaping an individual's decisions. For example, subjects systematically deviate in dictator and ultimatum games from normative predictions that are based on pure self-interest; results from bilateral bargaining experiments show that only a fraction of the subjects care about more than material payoffs to themselves.[44] The empirical results of Kahneman, Knetsch, and Thaler are particularly relevant in this regard.[45] They devised an environment wherein an individual may choose one of two

---

[39] *Id.* at 407–08.

[40] *Id.* at 408.

[41] *Id.* at 412.

[42] Bornstein & McCabe, *supra* note 5, at 464–65 (distinguishing between the incentives provided by common experimental economics research with those present in a real jury case, and suggesting that it makes such research less applicable to mock juries).

[43] *Id.* at 454.

[44] *See, e.g.*, Colin Camerer & Richard H. Thaler, *Anomalies: Ultimatums, Dictators, and Manners*, 9 J. ECON. PERSP. 209, 210–12, 216 (1995) (assessing the validity of the dictator and ultimatum games' non-self results and their deviation from game theory outcomes).

[45] *See* Daniel Kahneman, Jack L. Knetsch & Richard H. Thaler, *Fairness and the Assumptions of Economics*, 59 J. OF BUS. S285, S288 (1986) (analyzing how fairness effects individuals' decision-making choices using three different studies); *see also* Daniel Kahneman, Jack L. Knetsch & Richard Thaler, *Fairness as a Constraint on Profit Seeking: Entitlements in the Market*, 76 AM. ECON. REV. 728 (1986).

persons with whom to share a certain amount of money, knowing that one had made an unequal offer in a previously played ultimatum game while the other made an equal offer.[46] The payoff is such that the individual obtains more money if he or she chooses to share with the first person (unequal offer).[47] However, a majority chose to share with the second person (equal offer) even though it meant that the individuals also received less money.[48] The result suggests individuals are willing to sacrifice their personal gain to punish past unfair behavior, even when such behavior was not directed towards them.

Given the misleading nature of standard incentive mechanisms in the context of jury studies, it is critical to devise a mechanism that simulates the social utility of real juror experience. Without the proper alignment of decision-making incentives, an understanding of the true behavior of individuals in jury simulations may not be achieved.

II. AN INCENTIVE-ALIGNED MECHANISM FOR SOCIAL UTILITY

To create a mechanism that effectively replaces the missing elements of real-world decision-making, one must identify all of the components in play. A comprehensive analysis is necessary to accomplish this goal. In addition to the conventional components such as personal financial benefit, this analysis must explicitly incorporate the social utility an individual derives from making a correct judgment for third parties. Once the components are identified, it is possible to articulate a proxy incentive mechanism that could be used in mock jury research that will mimic the social utility experience by actual jurors.

A. Mathematical Framework of Juror Utility

A juror's utility has three major components. The first component is the disutility arising from to the effort (e_i) that an individual, i, must expend to fulfill jury duty. The effort includes the time to complete the decision-making task as well as additional burdens such as appearing at the courthouse, time away from work and/or family, etc. However, a juror may spend significantly more effort beyond the minimum if the problem is

[46] Kahneman, *supra* note 45, at S290–92.
[47] *Id.* at S290–91.
[48] *Id.* at S291.

complex, and the juror wishes to ensure that the best decision results.

The second component is the personal utility that can be derived from the decision. In other words, how the juror personally benefits. This can be further divided into direct and latent personal utility. Direct personal utility includes the cash payment or publicity a juror receives. Latent personal utility includes personal gains from the decisions made by the jury. For example, an individual without health insurance may derive latent benefits from a ruling that he or she perceives will better enable competitors to bring down the cost of pharmaceuticals. Alternatively, jurors may see latent benefits to a ruling that sends a message consistent with a social or political agenda. The most prominent example of this latter motivation is "jury nullification," wherein a jury appears to purposely disregard the law in reaching a verdict (often because it is perceived to be unjust).[49] Note that the degree of latent personal gain depends positively on the effort spent to achieve the desirable outcome, while the direct personal utility does not. One may define $u_{ip}(e_i)$ as individual i's personal utility, u_{ip}^{0} as the direct personal utility, and $u_{ip}^{1}(e_i)$ as the latent personal utility:

$$u_{ip}(e_i) = u_{ip}^{0} + u_{ip}^{1}(e_i) \quad (1)$$

The third component is the social utility derived from making the correct judgment for third parties. It can be further divided into two parts, which includes participation utility and justice utility. The participation utility is the utility an individual derives from the knowledge that he or she has fulfilled one's responsibility as a citizen. The justice utility depends on the welfare of the parties actually involved in the case; it reaches a maximum if the juror makes the correct decision, and becomes smaller the farther the decision is from the correct one. Note that the degree of justice utility depends positively on the effort spent to understand a case and make an educated decision, while participation utility does not. If one defines $u_{is}(e_i)$ as individual i's social utility, u_{is}^{0} as the participation utility, and $u_{is}^{1}(e_i)$ as the justice utility, the relationship is as follows:

[49] *See* Lars Noah, *Civil Jury Nullification*, 86 IOWA. L. Rev. 1601, 1604–05 (2001) ("Nullification occurs whenever a jury intentionally ignores the trial judge's instructions on the applicable law.").

$$u_{is}(e_i) = u_{is}^0 + u_{is}^1(e_i) \quad (2)$$

The utility function for individual i as a juror is thus:

$$\begin{aligned} U_i(e_i) &= u_{ip}(e_i) + u_{is}(e_i) - e_i \\ &= u_{ip}^0 + u_{ip}^1(e_i) + u_{is}^0 + u_{is}^1(e_i) - e_i \end{aligned} \quad (3)$$

In most cases, a juror will be unable to derive latent personal utility. In this case, the utility could be simplified as:

$$U_i(e_i) = u_{ip}^0 + u_{is}^0 + u_{is}^1(e_i) - e_i \quad (4)$$

B. *An Incentive-Aligned Mechanism for Implementing Social Utility*

While a real juror experiences the utility tradeoff as specified in equation (3), mock jurors who participate in jury simulation studies have quite a different utility function. Since the study is hypothetical and there is no real consequence to the juror's decision, a mock juror does not experience any social utility. Furthermore, he or she will also have no latent personal gain. As a result, the mock juror's utility is simply:

$$U_i(e_i) = u_{ip}^0 - e_i \quad (5)$$

In view of the missing incentive components, it seems clear that the subject's decision is likely to suffer from hypothetical bias. Most troubling is the fact that the missing utility components are functions of the effort, rather than a constant (which only help to determine whether to participate). As a result, a mock juror will likely spend effort that is substantially different in nature from that which he or she would spend if juror were participating in an actual case.[50] Because the missing incentive has positive utility, one expects that the mock juror will devote substantially less effort to decision-making.

To correct for the hypothetical bias, it is important that an incentive-aligned mechanism be implemented to give a mock juror additional utility that is consistent with that which would exist for a real jury. Of the three missing utility components

[50] MacCoun, *supra* note 8, at 513–14 (noting that the most straightforward consequence of hypothetical bias "is that real jurors may try harder than mock jurors—think harder, deliberate longer and ponder more deeply," but arguing that the available literature does not clearly establish this effect).

($u^1_{ip}(e_i)$, u^0_{is}, and $u^1_{is}(e_i)$), the first two can be readily remedied. The component $u^1_{ip}(e_i)$—arguably uncommon in most litigation—could be remedied by allowing an individual to receive additional personal compensation if the decision follows the pattern suggested by the self-interest bias under study. Similarly, component u^0_{is} could be added back to a mock juror's utility by paying that juror an additional fixed amount of money.

On the other hand, the third component, justice utility ($u^1_{is}(e_i)$), is the most critical and presents the biggest challenge for researchers and practitioners. While there are potentially many ways to implement such a corrective mechanism, we believe that three necessary conditions should be satisfied by all such mechanisms:

(1) The incentive mechanism must reflect social utility (or utility derived from rearranging welfare distribution among third parties) instead of personal utility;
(2) The incentive mechanism must reward the effort in a similar manner as it would in a real jury context; and
(3) It is important that the mechanism be flexible enough to accommodate heterogeneity across different types of cases.

With these conditions in mind, it is now possible to articulate a particular embodiment of the incentive-alignment mechanism outlined above.

C. *The Practical Application of an Incentive-Aligned Mechanism for Eliminating Bias*

One can imagine a generic incentive-alignment mechanism that is adaptable to a variety of legal cases in order to simulate actual juror utility. The participants are asked to apply unique and complex law and facts gathered substantially from the information provided during the study. Social utility is mimicked by establishing that a certain amount of money is at issue in the case and must be divided between two parties (the plaintiff receives what is required to compensate for the defendant's wrongful act, and the defendant keeps the rest). Participants should understand that is a legally correct division of money (the "reference decision"), and any deviation from that will unfairly punish one party while rewarding the other party undeservedly.

The proposed mechanism—explained to the mock jurors before a jury simulation study and imposed after they complete their verdicts—involves the following steps: (1) experimenters

identify two parties *a priori*, such that one party is generally considered to be deserving of a monetary gift (such as a well-respected charity like UNICEF), and the other party is generally considered to be undeserving of such largesse (such as a passerby in the street);[51] (2) upon completion of the experiments, the experimenters randomly pick a mock juror and compare his or her verdict to the reference decision (e.g., legally correct decision), and the percentage of deviation from the reference decision is calculated (e.g., 20%); and (3) experimenters then divide a certain amount of money (e.g., $200) such that the undesirable party (e.g., passerby) receives the total amount multiplied by the deviation (e.g., $200*20%=$40), and the desirable party (e.g., UNICEF) receives the rest ($140). This mechanism thus provides a mock juror with social utility, the magnitude of which (the amount of money the desirable party receives) is dependant on the mock juror's ability to deliberate the case in a legally correct manner.[52]

The mechanism is flexible and can be used to accommodate specific cases by changing the following three elements: (1) the two parties used, (2) the total amount of money to be divided, and (3) the relationship between the deviation and division of money.

III. EXPERIMENTAL RESULTS

To test the effectiveness of the proposed incentive structure, the authors conducted two experiments implementing the structure, and the results were compared to a control group under a conventional hypothetical condition. To more realistically demonstrate the application of the mechanism, the authors used a specific context. The experiment utilized a problem that is becoming quite common in high stakes mock jury analysis: an intellectual property infringement case in which a juror is required to assess the appropriate amount of money (damages) to award to the plaintiff (i.e., subsequent to a determination of liability). However, the authors concede that assessment of some essential aspects of jury behavior, such as

---

[51] These two parties do not correspond to the two parties involved in litigation (defendant and plaintiff), but rather, they represent two states of outcome (a desirable outcome and an undesirable outcome).

[52] Of course, a person who does not care about social utility will behave no differently when this mechanism is used, but this is quite realistic, as such a person would not take social utility into consideration if that person were a real juror.

deliberation,[53] were sacrificed in the interest of economy and clarity of results (though an approximation is employed). The methods should be equally effective if additional methodologies are added.[54] On the other hand, design problems that plague previous assessments of consequentiality in jury studies—namely, ensuring that participants believe the outcome of the consequential arm is "real"[55]—are far less an issue here because there is in fact no deception.

A. *Experiment One*

To achieve high external validity, realism in design was considered in concert with a format that would yield statistically significant and replicable results. The experimental context (legal case) and individual decision-making in a group environment were primary foci.

1. Experimental Design

A very realistic and complex legal case, loosely based on the facts of a real litigation that took place in the U.S. District Court for the District of Massachusetts,[56] was prepared for the study. It permitted the simulation of the conditions of a typical jury study to ensure that any observed effects could be generalized. It also ensured that subjects were not tested simply on abstract math skills by requiring subjects to grasp legal and factual aspects of the case as real jurors do.

The hypothetical case involved a patent on a method of producing large quantities of erythropoietin (a protein that stimulates red blood cell production) in tissue culture cells.[57] The

---

[53] *See* Bornstein & McCabe, *supra* note 5, at 459 (stating that "the lack of jury deliberations in experimental studies has been identified as one of the major threats to external validity and policy relevance").

[54] Such additions are probably required to firmly establish that *relevant* hypothetical bias is occurring and addressed by the proposed method. It is, however, beyond the scope of this initial work.

[55] Bornstein & McCabe, *supra* note 5, at 457–58.

[56] Amgen, Inc. v. Hoechst Marion Roussel, Inc., 126 F. Supp. 2d 69 (D. Mass. 2001). Note, however, the alleged acts of infringement described in our experiment were entirely fictional and should not be attributed to the either of the parties involved in that case. Because the case was more complicated and ended in a bench trial rather than a jury trial, the outcome is not directly comparable to our experimental results.

[57] All study instruments are on file with the authors and available upon request.

patent's owner sued a competitor for making use of erythropoietin production methods that allegedly infringed the plaintiff's patent. The study participants were given detailed facts surrounding the development and sale of the two companies' erythropoietin products as well as the law regarding patent infringement and damages.[58] The participants were informed that the liability phase of the case was already decided, with the defendant found to have infringed the plaintiff's patent, and the only remaining task was to decide the damages the defendant owed for the infringement over the time period involved in the case. The question of infringement damages was divided into six specific periods, each of which required an answer.

Determining damages in a case like this is often quite complex, hinging on such facts as a party's actual sales and whether the plaintiff would have been able to make additional sales if the defendant had not been on the market. The facts were written such that a single legally correct[59] answer existed for each of the six time periods, but could be derived only with significant mental effort. The key elements in each problem and the correct and most likely incorrect answers are summarized in Table 1.

[58] The statements of law were derived from model jury instructions that are used in actual trials. *See* AM. INTELLECTUAL PROP. LAW ASS'N, GUIDE TO MODEL PATENT JURY INSTRUCTIONS (1998), *available at* http://www.aipla.org/Content /ContentGroups/ Publications1/Guide_to_Model_Patent_Jury_Instructions.htm.

[59] The legally correct damages amount could be determined by applying the law to the facts contained in the test instrument.

Table 1. Experiment 1 Problems and Answers

Prob.	Legal/Factual Issue	Correct Answer	Answers from Common Mistakes
1	If the plaintiff can make no sales, a royalty is the sole measure of damages	$9,000	$90,000 (award plaintiff lost profits damages)
2	Plaintiff can obtain lost profits for defendant's sales within plaintiff's excess sales capacity; royalty for the rest	$495,000	$900,000 (award plaintiff lost profits for all sales) $90,000 (award royalty for all sales)
3	For lost profits damages, plaintiff must demonstrate by a preponderance of the evidence a reasonable probability of lost sales	$990,000	$1,800,000 (award plaintiff lost profits for all sales) $180,000 (award royalty for all sales)
4	Evidence sufficient to justify lost profits must be relevant	$1,350,000	$945,000 (misread facts and award royalty for 500 units) $1,800,000 (award plaintiff lost profits for irrelevant speculation)
5	A reasonable probability of lost sales cannot be defeated by the mere possibility of a different outcome	$1,800,000	$900,000 (accord improper weight to speculative argument) $2,700,000 (award plaintiff lost profits based on excess capacity instead of lost sales)
6	One cannot obtain any damages after the expiration of a patent, even if the defendant was unaware of the expiration	$1,080,000	$1,800,000 (award lost profits for period after patent expires) $990,000 (award lost profits for half sales instead of sales occurring in half year)

Actual jurors make decisions in a group environment in which they are exposed to each other's opinions while arriving at their own final determination. This is a key element of juror deliberation that the experiment intended to replicate in a carefully controlled manner.[60] The design involved providing study participants with the additional information one would be exposed to in a real jury; specifically, multiple (different) answers to a given question and the percentage of people in the group that adhere to each. This design allows for the precise control over the amount of information to which each participant is exposed. In the experiment, two or three possible answers are provided for the questions, each answer being associated with a percentage of prior jurors who have selected it. The various answers were obtained from the most common calculations observed in a pre-test set of experiments (e.g., percentages of subjects producing the correct answer and the most common mistakes, see Table 1).

Clearly, the use of such percentages presented a problem. Should the correct answer always be assigned to the highest percentage (i.e., treating it as the majority opinion)? If so, the simple, effortless, and commonly adopted strategy of agreeing-with-the-majority will necessarily lead to the correct answer and confound the hypothetical bias to be tested. One can assume that the outcome (how many individuals reach the correct answer) could be contaminated.[61] Conversely, the phenomenon of majority bias would not affect the outcome if the correct answer were not presented as the majority opinion. However, perceptive and careful subjects could become suspicious of the overall test design if "majority" answers are never correct. A creative solution was necessary to address this complication.

To rectify the majority bias issue, half of the questions from the case were selected and the majority answer was retained as correct (1, 2, and 4), while half of the responses were changed to reflect an incorrect majority opinion (3, 5, and 6). The obvious

---

[60] Alternate means of simulating group decision-making exist, but would be problematic for a proof of concept analysis. For example, one could assign individuals to a group and allow them to talk within their group before final deliberation. This has two major limitations: (1) it is hard to control the discussion and dynamics within each group; and (2) the group effectively becomes the unit of analysis, which makes it prohibitively expensive to obtain multiple data points in the experiment.

[61] Interestingly, it is hard to predict, *a priori*, who is more likely to adopt this strategy, a subject in the hypothetical condition or a real juror.

downside to this solution is that meaningful results from subjects' answers to the majority-correct questions could no longer be obtained. Therefore, the analysis and discussion must necessarily focus on the majority-incorrect questions.[62]

2. Experimental Procedure

Subjects were undergraduate business students at a large U.S. university. They randomly signed up for one of two sessions. A total of twenty-five (25) subjects showed up for the first session and were used for the hypothetical condition. After signing an informed consent form,[63] each subject was given a handout consisting of general instructions, a summary of facts and law involved in the case, and a special verdict form that requested answers for the aforementioned six time periods. The subjects were permitted to read and complete the special verdict form at their leisure. It took about twenty to forty minutes for all subjects to complete the study.

A total of twenty-eight (28) subjects appeared for the second session, which was used for the incentive condition. Before the start of the session, one subject was asked to go outside the classroom and find a random student walking by the building. That subject was to explain to the potential recruit that he or she would receive a minimum of $10 for coming to the classroom and remaining for about forty-five minutes to an hour (during which time that person could quietly do as he or she wished). The recruit was also informed that there was a possibility that he or she would receive substantially more money—up to $100—depending on the outcome of the experiment going on in the classroom. A female undergraduate student was eventually recruited and introduced to all subjects in the classroom; none of the subjects were acquainted with the recruit.

---

[62] Technically, on the jury forms used in the experiment, a subject had the option to accept a majority opinion even if it is different from his or her own calculated answer. In the experiments described below, however, a negligible number of participants acted in this fashion, and there were no significant differences between the hypothetical and incentive conditions. Thus, the data analysis presented in this paper provides only the actual answers each subject stated regardless of whether they chose to defer to the majority opinion or not. This helps to capture individual decision outcome, but the results are substantially unchanged if we replace these answers with the majority opinion.

[63] Approvals for human subject research were obtained by the authors for all experiments described herein. Documentation is available upon request.

The same experimental handouts used with the hypothetical group were then distributed to the subjects. They were asked to read the general instructions carefully, scan the summary and special verdict form, and then stop. The experimenter then put $100 in cash on the table in front of the group and described on the blackboard, lecture style, how their answer would be used to divide the $100:

> Two parties were designated, the first being the recruit in the room, the second being the Make-A-Wish foundation (introduced by showing the subjects the Foundation's website on a large projection screen which detailed its mission).[64] One of the subjects would be randomly selected at the end of the experiment, and one out of the six answers on that subject's verdict form would be randomly chosen. The subject's answer for that selected question would be used as the basis for dividing the $100 between the two parties (unless the subject marked an option that stated "I will agree with the majority even though my answer is different," in which case the majority answer would be used instead).

Subjects were shown by example how the experimenter would calculate the percentage of deviation (if any) between the subject's answer and the correct answer: If the subject's answer is $2500, and the correct answer is $2000, the deviation is (2500-2000)/2000=25%. The result would be the same if the subject's answer were 25% below the correct answer. The deviation is bounded by 100% more than the correct answer; any answers above that will be treated as 100% deviation (e.g., a $10000 answer has an effective deviation of 100% even though (10000-2000)/2000=400%). The product between the $100 cash and the percentage of deviation would be the amount of money presented to the student recruit, and the remainder would be sent to the Make-A-Wish Foundation. In the preceding example, $25 would go to student recruit and $75 to the Make-A-Wish Foundation. Additional examples were given where the subjects were asked to figure out the resulting division to ensure they understood the incentive.

---

[64] The Make-A-Wish Foundation is a well-known national charity that grants the last wishes of terminally ill children. The authors are not aware of any political or religious affiliation on the part of the Foundation.

3. Results

It was predicted that subjects under the incentive conditions would behave differently, even though they do not benefit personally from the outcome. A direct prediction of the result of this enhanced effort is that more of these subjects would obtain the correct answer. To verify this prediction, the responses to the three questions that were free of the confounding effects discussed above were studied.

To examine the performance across questions, the number of subjects with the same answers was counted, and the numbers tabulated into four groups corresponding to either one of the three presented answers or all other answers that did not match any of the presented ones (see Table 2).

Table 2. Cross Question Results in Experiment 1[65]

	Quest. 3		Quest. 5		Quest. 6	
	Hypo.	Incent.	Hypo.	Incent.	Hypo.	Incent.
Majority Opinion	9 (36%)	10 (36%)	10 (40%)	6 (21%)	6 (24%)	6 (21%)
2nd Common Opinion	10 (40%)	10 (36%)	9 (36%)	18 (64%)	11 (44%)	17 (61%)
3rd Common Opinion	3 (12%)	2 (7%)	3 (12%)	1 (4%)	6 (24%)	3 (11%)
Other Opinions	3 (12%)	6 (21%)	3 (12%)	3 (11%)	2 (8%)	2 (7%)
Total	25 (100%)	28 (100%)	25 (100%)	28 (100%)	25 (100%)	28 (100%)

As predicted, subjects in the incentive condition performed better for questions 5 and 6. In question 5, 64% of subjects in incentive conditions obtained the correct answer, compared to 36% in the hypothetical condition, and the difference is statistically significant based on the chi-square test (p=0.00).[66]

[65] The second common opinion is the correct answer for all three questions.

[66] We used the percentage in the hypothetical condition as the expected percentage for the incentive-aligned condition.

For question 6, 61% of subjects in the incentive condition were correct, compared to 44% in the hypothetical condition, and this difference is significant (p=0.07). These findings provide strong evidence that individuals indeed act differently under the proposed incentive structure as compared to the pure hypothetical condition. An additional interesting observation is that for question 5, but not for question 6, the number of subjects that have the same opinion as the majority is significantly higher in the hypothetical condition compared to the incentive condition (p=0.04).

Unfortunately, the results for question 3 appeared to suffer from a problem in the construction of the instrument. Following the experiment, subjects were interviewed and it was discovered that the question could be interpreted in more than one way. An answer designated as wrong—the majority answer—could in fact be considered legally correct if the question were read in a particular light. Subjects putting in the most effort may have chosen an incorrect answer. Therefore, responses for this question are ambiguous and appropriately discarded.

Next, examined performance at the individual level was studied. The number of the questions in which a subject provided the correct answer was counted (0, 1, 2, or 3). Given our assessment that question 3 was ambiguous, we also examined the result for questions 5 and 6 only (0, 1 or 2). Both results are included in Table 3.

Table 3. Individual Performance in Experiment 1

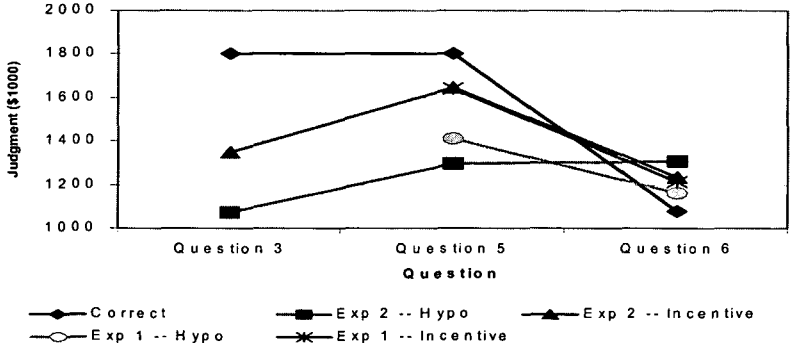
Number of Questions Correctly Answered	Quest. 3, 5, 6		Quest. 5, 6	
	Hypo.	Incent.	Hypo.	Incent.
0	7 (28%)	4 (14%)	10 (40%)	5 (18%)
1	9 (36%)	7 (25%)	10 (40%)	11 (39%)
2	7 (28%)	13 (46%)	5 (20%)	12 (43%)
3	2 (8%)	4 (14%)	N/A	N/A

It is clear that subjects in the incentive condition performed better in terms of total correct answers, and the difference is statistically significant based on chi-square test (p=0.06). The same observation holds when question 3 was removed (p=0.00). These results suggest that the incentive-aligned group appeared to have a better grasp of the facts and law presented in the summary and worked harder to obtain the correct result.

Lastly, the average answer of question 5 and question 6 was calculated and studied. The average is a useful measure because it permits one to determine whether the group as a whole has moved toward a more accurate result, or if the numbers of correct or near correct answers are washed out by the magnitude of incorrect ones. If one were attempting to obtain an overall picture of the group response—such as determining a likely jury verdict for a given group of individuals—one would wish to know whether the results could be different in the incentive-aligned group. Because the amount of deviation would not necessarily be a function of effort, the outcome seems inherently unpredictable. It could differ from case to case, and within a case it may depend on the specific question asked.

We plotted (see Figure 1), for questions 5 and 6, the correct answer, the average of those in the incentive condition and those in the hypothetical condition.[67] As suspected, results for questions 5 and 6 are not the same.

Figure 1. Average Answers for Each Question



[67] We omitted Question 3 here for two reasons. First, the question is ambiguous as discussed above, and second, we plotted the results for both experiment 1 and experiment 2 in the same graph for ease of comparison, and Question 3 in experiment 2 has a different correct answer

The average in the incentive condition is statistically smaller than the correct answer (p=0.02), and it is marginally statistically higher than the average in the hypothetical condition (p=0.08). For question 6, however, the averages in the incentive and hypothetical conditions are not statistically different (p=0.31).

While experiment 1 provides strong support for the general predictions in this paper and the validity of the incentive structure, it does have a major limitation. The impact of the third-party players could be inconsistent. In general, the Make-A-Wish Foundation appears to be an appropriate party to inspire largess because it is almost universally acceptable to potential subjects and accessible to future experimenters and practitioners. However, the other party—a randomly selected passerby—could provoke different responses from the participants depending on his or her identity. For example (although there is no evidence it impacted Experiment 1), it is possible that an unusually physically desirable or needy passerby may make subjects indifferent between dividing money between the Foundation and the passerby. With this issue in mind, the nature of the third parties and other aspects of the experiment were modified for the next round of studies.

B. Experiment Two

Experiment 2 was undertaken with three methodology goals in mind: the replication of the positive results from experiment 1, the implementation of a remedy for the ambiguous nature of experiment 1's Question 3, and the use of a more objective party than the random passerby. Additionally, the underlying reasons for the differences between the hypothetical and incentive-aligned group were sought. Given the fact that the incentive mechanism rewards social utility, it was predicted that a bigger difference would be observed for those subjects who care more about social welfare. The analysis here focuses on Questions 3, 5, and 6.[68]

1. Experimental Design and Procedure

The instrument used in experiment 2 was the same as in

[68] The results for the other three questions have been assessed and are available from the authors upon request.

experiment 1, except for the following: (1) question 3 was modified to eliminate the ambiguity;[69] (2) a new embodiment of the incentive structure was implemented; and (3) a brief survey to understand subjects' preference for socially desirable behavior was presented at the conclusion of the study.

Again, subjects were undergraduate business students recruited from a large U.S. university approximately two months after the completion of experiment 1 (none of them participated in experiment 1). They randomly signed up for one of two sessions, each of which was further split into two groups. To ensure that there was no difference between the two groups in each session, all subjects were first placed in a single room and half of the subjects were randomly selected and then moved to a different room. One room was used for the hypothetical condition and one for the incentive condition. There were a total of thirty-three (33) subjects in the hypothetical condition and thirty-two (32) subjects in the incentive condition.

For the incentive condition, the subjects were informed as to how $100 would be divided between two parties as in experiment 1 except that the counter party to the Make-A-Wish Foundation was different. The subjects were told instead that any deviation of the randomly selected result from the correct answer would result in the corresponding money being completely wasted. Specifically, the money would be used to purchase expensive coffee (or coffee beans) from a Starbucks café on campus, which would then be promptly dumped down a drain (coffee) or into a trash can (coffee beans). The subjects were told that this would be done immediately following the experiment, and they were invited to stay and watch. In the experiment, the randomly selected answer in session 1 resulted in an $83 deviation. Two subjects were paid $5 to go to the café afterwards and purchased $83 worth of coffee beans that were then dumped into the trashcan in the classroom upon their return. The random answer in session 2 was 100% correct, thus all $100 went to the Make-A-Wish Foundation.

[69] To obtain the correct answer, a participant must derive from the study instrument the legal point that a reasonable probability of lost sales cannot be defeated by the mere possibility of a different outcome (compare to Table 1). The correct answer is $1,800,000, and two likely incorrect answers are $990,000 (accord improper weight to speculative argument) and $180,000 (award royalty for all sales).

After each subject completed his or her deliberation, the subject was asked to take a short survey after turning in the verdict form.[70] The short survey was intended to measure social desirability by employing "[a] summated ratings scale purporting to measure the degree to which people describe themselves in socially acceptable terms to gain the approval of others."[71] It was originally developed by Crowne and Marlowe[72] and has been used extensively in the behavioral literature.[73] The total scores range from zero to thirty-three, and a person with higher score tends to respond to questions in a socially desirable manner, whereas a person with lower score is less likely to answer the question that way.

2. Results

Similar to experiment 1, the aggregate outcome for questions 3, 5 and 6 by problem was examined first (Table 4). Consistent with the prior results, 59% of the participants were observed to have chosen the correct answer in the incentive condition for question 5 versus 36% in the hypothetical condition. This difference is statistically significant (p=0.01). The number of people who selected the majority answer was statistically smaller in the incentive-aligned condition as compared to the hypothetical condition (p=0.01).

---

[70] On file with the authors, and available upon request.

[71] GORDON C. BRUNER II ET AL., 3 MARKETING SCALES HANDBOOK: A COMPILATION OF MULTI-ITEM MEASURES 616 (2001).

[72] *See* Douglas P. Crowne & David Marlowe, *A New Scale of Social Desirability Independent of Psychopathology*, 24 J. CONSULTING PSYCH. 349, 350–51 (1960) (describing how the authors developed the social desirability scale).

[73] *See, e.g.*, David Glen Mick, *Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism*, 23 J. CONSUMER RES. 106, 107 (1996) (explaining that the Marlowe-Crowne Scale is the most popular socially desirable responding scale); Richard G. Netemeyer et al., *Trait Aspects of Vanity: Measurement and Relevance to Consumer Behavior*, 21 J. CONSUMER RES. 612, 617 (1995) (stating that when measuring vanity traits "an often ignored aspect of scale development is testing for social-desirability bias"; Marsha L. Richins & Scott Dawson, *A Consumer Values Orientation for Materialsim and Its Measurement: Scale Development and Validation*, 19 J. CONSUMER RES. 303, 310 (1992) (utilizing the social desirability scale to determine whether materialism is susceptible to the social desirability bias).

Table 4. Cross Question Results in Experiment 2[74]

	Quest. 3		Quest. 5		Quest. 6	
	Hypo.	Incent.	Hypo.	Incent.	Hypo.	Incent.
Majority Opinion	14 (42%)	14 (44%)	17 (52%)	9 (28%)	13 (39%)	6 (19%)
2nd Common Opinion[75]	10 (30%)	15 (47%)	12 (36%)	19 (59%)	14 (42%)	18 (56%)
3rd Common Opinion	4 (12%)	1 (3%)	0 (0%)	3 (9%)	3 (9%)	4 (13%)
Other Opinions	5 (15%)	2 (6%)	4 (12%)	1 (3%)	3 (9%)	4 (13%)
Total	33 (100%)	32 (100%)	33 (100%)	32 (100%)	33 (100%)	32 (100%)

For question 6, it was 56% versus 42%, however, unlike experiment 1, the difference was not statistically significant (p=0.11). On the other hand, the number of people who selected the majority answer is statistically smaller in the incentive-aligned condition as compared to the hypothetical condition (p=0.02). Interestingly, the percentage of subjects who obtained the correct answers under either condition was very close in both experiments. For question 5 in experiment 1, the percentage of correct answers was 64% and 36% for incentive and hypothetical conditions, respectively. For question 6 in experiment 1, it is 61% and 44%, respectively. This high consistency provides another piece of evidence for the robustness of this paper's hypothesis and experimental design. It is also worth noting that this consistency also holds true for those who have selected the majority opinion (which is wrong in these questions), but only for those in the incentive condition. To a large extent, this supports the notion that results in the incentive condition are more consistent across different runs of the same experiment (with different subjects), while the hypothetical condition subjects are driven by various home grown preferences, and their actions may be less consistent when they did not obtain correct answers.

[74] Once again, the second common opinion is the correct answer for all three questions.

[75] Once again, the second common opinion is the correct answer for all three questions.

As predicted, after we revised question 3 based on subject feedback, we found the percentage of subjects who obtained the correct answer to be significantly higher in the incentive condition (47%) as compared to that in the hypothetical condition (30%) (p=0.04). This result confirms the observations for questions 5 and 6 from both experiments.

When we examined the data at the individual level, consistent with Experiment 1, we found significant improvement of performance (p=0.00).

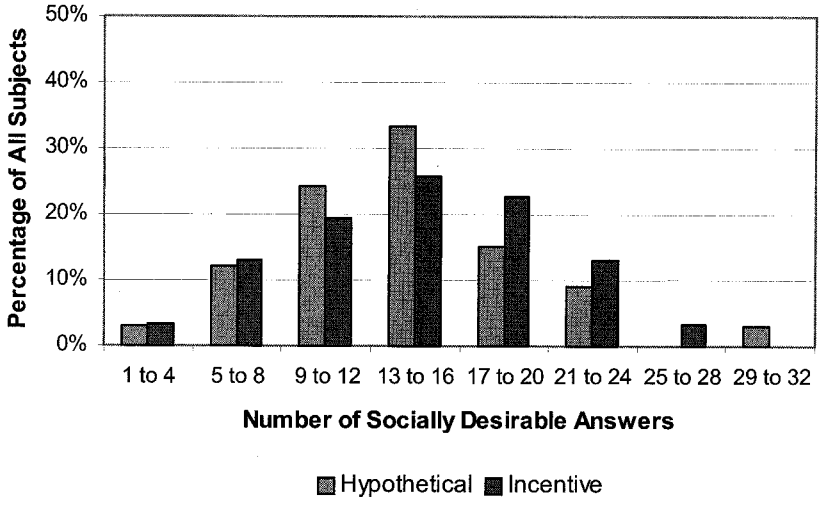
Table 5. Individual Performance in Experiment 2

Number of Questions Correctly Answered	Questions 3, 5, 6	
	Hypothetical	Incentive
0	9 (27%)	6 (19%)
1	15 (45%)	7 (22%)
2	6 (18%)	12 (38%)
3	3 (9%)	7 (22%)

The averages of the answers for each question are plotted in the same figure as those from experiment 1 for ease of cross-experiment comparison (Figure 1). Remarkably, the averages in the incentive condition are very similar between the two experiments for questions 5 and 6, respectively. Again, this provides additional evidence of the robustness of the incentive structure. For question 5, the average in the incentive condition is significantly lower than the correct answer (p=0.04), but significantly higher than the average in the hypothetical condition (p=0.00). This result is consistent with that obtained in experiment 1. For question 6, the averages in the incentive and hypothetical are not statistically different from each other (p=0.68). For question 3, the average in the incentive condition is significantly lower than the correct answer (p=0.00), but significantly higher than the average in the hypothetical condition (p=0.02).

Finally, the hypothesis that the individuals on whom the incentive structure has the greatest effect are those who care a great deal about third-party utility and doing the right thing for others was investigated. The total scores for subjects were calculated on the social desirability scale (ranging from one to thirty-three) for each subject and we plotted the distributions of these scores for both hypothetical and incentive conditions (Figure 2).

Figure 2. Distribution of Socially Desirable Answers for Each Subject



The average score across subjects was 13.9 with standard deviation of 5.7 in the hypothetical condition, and 14.5 for the incentive condition with standard deviation of 5.5. Although it may appear that the incentive condition has more subjects who exhibit social desirability, the average of the scores are not statistically different (p=0.35). Thus, the measures of social desirability also serve to provide evidence that samples for both conditions are equivalent. Since the distribution of scores in both conditions is close to normal with symmetric tails on both sides, the subjects were divided into three groups: (1) those (roughly) within one standard deviation of the average, (2) those (roughly) above one standard deviation from the average, and (3) those (roughly) below one standard deviation from the average. The

individual level of performance (how many questions a subject has answered correctly) was then calculated for subjects within each of the three groups, for both hypothetical condition and incentive condition (Table 6).

Table 6. Cross Individual Performance Segmented Based on Preference of social desirability in Experiment 2, Questions 3, 5, 6

Social Desirability[76]	Hypothetical[77]	Incentive
Low	1.17 (0.97)	1.83 (2.16)
Medium	1.13 (0.94)	1.5 (0.97)
High	0.75 (0.25)	1.71[78] (0.9)

While the average performance in the incentive condition is better than that in the hypothetical condition in all three groups, the only group where such an increase is significant is for subjects in the high social desirability group (1.71 for the incentive condition, compared to 0.75 in the hypothetical condition). This result supports the hypothesis that, if the incentive condition makes people work harder, which in turn leads to a higher probability of correct answer, the difference within the hypothetical groups is expected to be most significant for individuals with high social desirability values.

DISCUSSION AND CONCLUSION

The economic behavior of participants in studies intended to simulate contexts wherein a person other than the participant is affected by the outcome is a relatively unexplored field. The experiments herein demonstrate that such behavior can be significantly influenced by the presence of an incentive-

[76] The subjects are grouped into three segments, Low (those below 1 SD of the mean (scores 1-9)), Medium (those within 1 SD of the mean (scores 10-19)), and High (those above 1 SD of the mean (scores 20-30)).

[77] Average, with standard deviation in parenthesis.

[78] The average is significantly higher in the incentive condition compared to the hypothetical condition, among those with high social desirability (p=0.028).

alignment mechanism. Such a mechanism can cause participants to perform more accurately, suggesting that they commit more time and effort. The fact that the incentives are similar in nature to those experienced in real world contexts creates a strong presumption that their use can make simulations more realistic and useful.

The results are likely to have particular importance in the context of jury simulations. Americans rely on the results of such studies to determine how to best reform the judicial system.[79] Questions such as what types of trial procedures may evoke prejudice or to what degree do jurors benefit from note taking are paramount for both real and perceived equity. These inquiries are substantially informed by simulations, and it is essential that the accuracy of this work be raised to the highest level possible. Moreover, the business world is dramatically affected by litigation; knowing when to settle or pursue one's case in court often hinges on predictable jury information,[80] a product of simulations. This paper suggests that the use of the incentive-alignment mechanism detailed herein may be essential to obtaining the most accurate results from studies concerning groups like juries. Although ignoring the hypothetical bias occurring in the absence of the mechanism may not pose significant problems in every case, it is impossible to know for sure. In general, the use of such incentives should be a no-lose situation.

It is worth imparting some words of caution regarding the meaning and transferability of the results. First, although the incentive-alignment mechanism described herein is in the same category as the incentive experienced by actual jurors, and it has been demonstrated that the mechanism does indeed increase subject performance, one cannot state conclusively that this performance is actually closer to that of an actual jury. While it is logical to at least make the assumption that real juries are more accurate than hypothetical juries, they may be *less* accurate than subjects using our mechanism, particularly if the incentive is too strong. Unfortunately, this complication is highly fact

---

[79] *See supra* notes 3–4 and accompanying text.

[80] *See* Cahoy & Ding, *supra* note 3, at 36 (asserting that the substantial risk often attached to jury trials is especially "unacceptable in the context of financially-significant cases, and the urge to avoid the risk—by early settlement on undesirable terms, if necessary—is great").

specific, and would be impossible to prove conclusively without either (1) conducting several studies in the context of on-going litigations in the hope that one of the cases studied happened to be one of the very few litigations that ended in a jury trial as opposed to settlement or (2) reverse-engineering a realistic jury study starting with a decided case. The former would be prohibitively expensive and the latter would be exceedingly difficult to conduct in an objective manner, knowing the outcome of the litigation. To ensure that such an aberration does not occur, the authors suggest that future experimenters choose an incentive that is less strong than that experienced by an actual jury (though necessarily stronger than the hypothetical).

Second, it must be acknowledged that other aspects of jury behavior may blunt the effect of hypothetical bias. For example, the dynamics of group decision-making could have the effect of reducing (or amplifying) the consequentiality effect. Additionally, the particular question at issue in a study (e.g., damages, racial bias, etc.) may be more or less subject to hypothetical bias. Therefore, the authors do not argue that the above results render all mock jury research inherently invalid or unreliable. Moreover, when contributing behavioral factors are not accounted for, the described mechanism will likely not provide complete verisimilitude. This article merely suggests that non-self economic incentives may provide an important new technique for addressing a seemingly intractable research problem. The authors hope that this work will be carried on and expanded upon in many contexts to demonstrate its impact.