Fairness and the Willingness to Accept Plea Bargain Offers

Avishalom Tor, Oren Gazal-Ayal, and Stephen M. Garcia

In contrast with the typical assumption in plea bargaining law and economics, we show defendants may reject plea offers based on fairness considerations. Specifically, offers where the sanction clearly appears excessive for the crime ("substantively unfair") and offers that appear inferior to those received by others in similar cases ("comparatively unfair") diminish defendants' willingness to accept plea offers (WTAP). Part 1 analyzes real-world data in Study 1 and reviews early experiments, all of which suggest substantive fairness impacts WTAP but do not control for important confounds. Part 2 therefore presents Studies 2-4 that confirm the independent impact of substantive fairness. Part 3 then explores the evidence on comparative evaluations of fairness and presents Study 5 that confirms their independent effect on WTAP. Study 6 further shows substantive and comparative fairness can impact WTAP simultaneously and independently. Part 4 concludes, showing our findings require both supporters and detractors of plea bargaining to reexamine their positions.

The highly puzzling behavior of Paul Hayes, the small-time criminal defendant in Bordenkircher v. Hayes (434 U.S. 357 [1978]), highlights the significance of the studies reported here for the broader plea bargaining literature. Hayes was charged in 1973 for “uttering a forged instrument,” a crime that carried a sentence of two to ten years in prison. He stole a check in Lexington, Kentucky, made it out to a local grocer for $88.30, and forged the signature on it. The case was simple, the evidence solid, and both Hayes and the Kentucky prosecutor Bagby wanted to plea bargain. In fact, a plea bargain is the typical means by which the criminal justice system disposes of such simple cases.

Yet, Bagby's plea offer was all but typical. He offered Hayes a five-year sentence, which was very harsh for such a minor offence. The offer was backed up with a
powerful threat: If he were to refuse the plea, Hayes would be charged under Kentucky’s habitual criminal statute that mandated life in prison (because of two earlier felony convictions) in the highly likely case of his forgery conviction.

Bagby probably expected Hayes to enter a guilty plea. After all, the vast majority of criminal defendants plead guilty, (Wasik, Thomas and Redmayne 1999; U.S. Department of Justice 2004), mostly following a plea bargain (Bibas 2006). Hayes became famous, however, for rejecting the five-year plea offer. His rejection was especially striking in the face of a near-certain conviction at trial and his knowledge that he qualified for a mandatory life sentence. Following the rejection, Bagby carried out his threat and Hayes was sentenced to life imprisonment.\(^3\)

The behavior of Paul Hayes seems to defy common sense. Why choose a near-certainty of life in prison over the absolutely certain, but dramatically shorter, 5-year sentence offered in the plea? In fact, Hayes must have known (and, at any rate, his lawyers must have told him) that Bagby had to carry out his threat. In the circumstances, therefore, there seemed to be no logical reason for Hayes's plea rejection.

We argue that a better understanding of the reasons for Hayes's behavior would inform the debate on the desirability and appropriate scope of plea bargaining (e.g., Alschuler 1981; Schulhofer 1980, 1992; Scott and Stuntz 1992a). From a traditional law and economics perspective, Hayes's behavior appears plainly irrational: Plea bargaining defendants are assumed to reveal a preference for the discounted sanction offered in the plea over the expected sanction at trial (e.g., Bar-Gill and Gazal-Ayal 2006; Easterbrook 1983, 1992; Landes 1971; Scott and Stuntz 1992a). Similarly, defendants who refuse a plea are assumed to reveal a rational preference for the risky trial over the discounted plea. A rational 29-year old defendant like Hayes, however, would never exhibit the dramatic risk-seeking revealed by choosing an almost certain life sentence at trial over the deeply-discounted 5-year plea offer.

Hayes poses a puzzle, moreover, even for those legal scholars who criticize the rational defendant view and assert, instead, that defendants seek to minimize their sanction but are often unable to do so. These scholars explain that defendants are led to accept inferior plea offers because they are subject to pressure, lack of information, biases and more (Alschuler 1975; Schulhofer 1988; Bibas 2004). Hayes, alas, did exactly the opposite of what these scholars expect, rejecting what he knew and understood to be an exceptionally attractive offer in the shadow of a life sentence at trial.

We would like to suggest that Hayes's puzzling behavior may have been driven by factors that commonly affect defendants' subjective evaluation of the fairness of plea terms and, consequently, their willingness to accept plea offers. Specifically, Hayes may well have thought the 5-year offer excessive for his crime and thus "substantively unfair." He also seems to have found Bagby's threat of a life sentence for refusing to accept the plea offer dramatically inconsistent with the offers made to other criminal defendants in similar cases (Stuntz 2005), violating his "comparative fairness" notions.

---

\(^3\) The sentence was upheld by all the state courts and the lower federal court. It was overturned by the Federal Court of Appeals for the Sixth Circuit, only to be reinstated by a Supreme Court majority.
In this article we show that both of these fairness variables significantly affect defendants’ willingness to accept plea offers (WTAP). Part 1 begins by examining the extant evidence on the likely impact of substantive fairness on WTAP. It shows how anecdotal evidence, our analysis of real-world data in Study 1, and two early experimental studies all suggest the importance of this variable. None of these sources, however, controls for potential confounds, most notably the impact of expected sanction on WTAP. We therefore present in Part 2 Studies 2-4 that use controlled hypothetical scenarios to confirm the independent impact of substantive fairness on WTAP and place it within a law-and-economics framework. Importantly, moreover, these studies also relate substantive fairness to the broader behavioral decision making framework, revealing its interaction with egocentric bias as well as its central role in shaping defendants’ plea bargaining risk preferences. Part 3 then explores the evidence on the impact of comparative evaluations of fairness and presents study 5 that confirms their independent effect on WTAP. Study 6 further shows that substantive and comparative fairness considerations can impact WTAP simultaneously and independently of one another. Part 4 concludes by examining the implications of our new findings for the plea bargaining debate. In this part we show how our empirical evidence reveals new challenges and opportunities for plea bargaining that require both its supporters and its detractors to reexamine many of their long-standing positions.

1. EVIDENCE ON SUBSTANTIVE FAIRNESS IN PLEA ACCEPTANCE

The account Paul Hayes provided in his Appeal to Petition for Certiorari may provide tentative insight into the behavioral causes of his surprising refusal to plead guilty. Hayes appears to have thought the sentence offered was too harsh for his small-time forgery (Stuntz 2005) – that is, substantively unfair. More generally, an offer may violate defendants’ substantive fairness notions when it seeks to impose a sanction they find clearly inappropriate for the offense. In the extreme case of an innocent defendant, for instance, any sanction demanded in a plea offer would likely appear substantively unfair.

The following sections therefore review the significant body of real-world and early experimental evidence on substantive fairness and pleas. For simplicity, we focus on the most obvious case where a plea bargain is substantively unfair – namely, where the defendant is innocent. We begin by citing some anecdotal evidence, and augment it with our empirical analysis of real-world exoneration data and corroborating findings from past experimental tests. This analysis shows that although the evidence strongly suggests substantive fairness impacts WTAP, important questions remain: The anecdotal evidence on seemingly irrational plea rejections, while striking, may be rare cases of extreme judgmental error, involving innocents who had private information about their conviction probability, or reflecting optimistically biased judgments of innocents rather than their fairness-driven preferences. The concerns over such confounds similarly

---

4 Wrongful conviction is often referred to as unfair or immoral in philosophy (e.g., Feinberg 1974), law (e.g., Dworkin 1985; Packer 1968, p. 250), psychology (e.g., Piaget 1932, p. 197), and everyday usage (e.g., Oxford English Dictionary). Note also that the intriguing question of the impact of the requirement to admit guilt per se – as distinct from the effect of the sanction – is beyond the scope of the present analysis.
plague our empirical analysis. Even the two early plea bargaining experiments, which used simulations and elaborate hypothetical scenarios, may have confounded substantive fairness with either probability of conviction beliefs or the expected sanction at trial.

We therefore complement this evidence with simple experimental scenarios of our own that control for the potential confounds, including that of the expected sanction at trial, and confirm the independent impact of substantive fairness on WTAP.

1.1. Real-World Evidence

A range of real-world evidence suggests that innocent defendants are less likely to accept plea offers, much like Hayes refused to accept what he felt was an excessive prison sentence that Bagby offered him. One illustrative case is Kelly Jarrett's, who rejected a plea bargain for time served on moral grounds, knowing that she was very likely to serve decades more in jail because of her refusal.5

The Tulia scandal, which concerned multiple defendants, similarly reveals the reluctance of innocents to accept seemingly attractive plea offers, even in the shadow of high expected sentences at trial. In 1999, forty three defendants were arrested for selling powder cocaine to an undercover investigator in Tulia, Texas, and all but one were charged. Four of these defendants, who had very strong alibis, achieved charge dismissals. The remaining thirty eight innocent defendants had to choose between trial and plea, knowing their chances of disproving a deputy sheriff's testimony were slim.

The first two defendants refused their plea offers, pleaded "not guilty" at trial, and were sentenced to 99 and 434 years of imprisonment. Nevertheless, the six defendants who followed still refused their plea offers. At their trials, the six received sentences of 12, 20, 25, 40, 45 and 60 years of imprisonment, respectively. Only following these outcomes, most of the remaining defendants (27 of 30) agreed to plead guilty. Their plea bargains usually resulted in probation or other non-incarceration sentences.6

Striking anecdotes such as Jarrett's case and the Tulia scandal do suggest that innocent defendants are reluctant to accept even those plea offers that appear objectively attractive alternatives to their expected sanction at trial. However, further to confirm the real world impact of substantive fairness on WTAP, we turned in Study 1 to a more systematic analysis of evidence on the behavior of innocents who were convicted at trial but later exonerated.

Study 1 focused on Higgins's (2003) database, which contains details of 316 known cases of wrongful felony convictions in the United States and Canada. This database was selected for two reasons: First, it contains sufficiently detailed information

---

5 Jarrett was convicted of a murder based on unreliable eye witness identification by an elderly man. After serving about 10 of her 25 years to life sentence, she refused an offer to plead guilty and be set free. A few years later, Claudia Angelos, an NYU law professor, succeeded in a habeas corpus challenge on Jarrett's behalf, and the district court ordered a retrial. Yet Angelos knew this decision is likely to be overturned on appeal. The state offered Jarrett to plead guilty for time served, but contrary to Angelos' advice she refused on moral grounds. On appeal, the circuit court overturned the retrial decision as expected and Jarrett served additional nineteen years until, after the story was featured by PBS, she was paroled in 2005. See http://www.pbs.org/wgbh/pages/frontline/shows/plea/four/jarrett.html.

6 The data on the Tulia defendants were collected from Sherrer (2003) and Amarillo Globe News (2001). Additional accounts of the story can be found in Gold (2003) and Gross et al. (2005).
on the different cases, including the method of conviction. This detail, in turn, allowed us to determine, for each exoneration case, whether these innocents chose plea or trial. Second, the Higgins (2003) database employs strict criteria for case inclusion, including only convicts who were released or pardoned by an official representing the government because new, post-conviction, evidence established the person's innocence.  

We categorized the records in the database on the variables of type of felony (mostly rape or murder), year of the wrongful conviction, type of conviction (plea or trial), causes for the mistaken conviction, and type of exonerating evidence.

An analysis of the categorized data revealed that only 19 of 316 (6.0%) of the exonerated were convicted following a guilty plea; in the remaining 94% of the cases, an erroneous jury decision led to the conviction. This 6% rate stands in sharp contrast to the common rate of guilty plea in felony cases; throughout most of the twentieth century guilty pleas accounted for 75% to 90% of felony convictions (Alschuler 1979).  

In fact, the 99% confidence interval around the 6% rate in our sample data has a lower bound of 2.54% and an upper bound of 9.48%. The dramatic difference between this confidence interval and the general felony cases guilty plea rate further emphasizes the strong tendency of innocents to opt for a trial notwithstanding conviction odds.

We also analyzed each of the common offenses in the database separately to get a more nuanced perspective on the data. Although a simple comparison of the raw plea rates for known wrongful convictions versus convictions in general makes the differences between the two groups readily apparent, we conducted further statistical tests to underscore these dramatic differences. However, our database spans over a century of cases, while record-keeping for U.S. Department of Justice annual felony statistics (published through the Bureau of Justice Statistics) is relatively new. We therefore used as a benchmark year for comparison purposes the earliest record we were able to obtain, which was 1990. The data for 1990 where then compared to a subset \( (N=134) \) of the 316 wrongful conviction cases in the database. These were the cases whose conviction years spanned the period of 1980 to 2000, setting the middle of the data subset period to 1990.

Echoing the general pattern found above, only 7% (9 of 134) of the 1980-2000 subset entered guilty pleas, compared to the 91% plea rate among all convicted defendants of violent crimes in the benchmark year 1990 \( (\chi^2=929.44, \ p<.00001) \). Strikingly, among the 66 rape and sexual assault exonerations during this timeframe, none (0%) followed a guilty plea; this figure is obviously significantly different from the 95% guilty plea rate among rape and assault convictions in 1990. Among murder and manslaughter cases we found a substantial 14% guilty pleas among exonerated

---

7 Cases in which the person received a new trial and was found "not guilty" are not included unless the exonerating evidence is DNA. Furthermore, cases were not included unless a written document describing the details of the case and exoneration was available.

8 Almost all the cases in the database refer to trials in the twentieth century. Importantly, guilty plea rates remain at a comparably low 5.7% (10 pleas of 175 cases) even when examining only post-1970 cases. Furthermore, Gross et al. (2005) obtained a similar guilty plea rate upon examining 340 exonerations which took place between 1989 and 2003. They found 20 defendants of the 340 wrongfully convicted defendants (5.9%) pleaded guilty. Thus, this 6% rate of guilty plea seems to represent old and new exoneration cases alike.

9 Intriguingly, almost all of the exonerations following a guilty plea in the database (17 of 19) followed a false confession to the police), a variable whose impact on WTAP is beyond the scope of this article.
defendants (9 of 66). However, the difference between this 14% rate and the 90% benchmark guilty plea rate among comparable convictions is still dramatic and highly significant ($\chi^2 = 138.89, p < .00001$). The relatively high number of guilty pleas in these cases may have resulted from the risk of a death sentence at trial; in fact, at least 6 of these 9 defendants explained their guilty plea was driven by their fear of capital punishment. It is therefore possible that when bargaining is conducted in the shadow of the death sentence, even innocents are more likely to plead guilty.

Importantly, moreover, the data indicates the rarity of wrongful convictions following a guilty plea does not result from the diminished tendency of such defendants to challenge their conviction, as some have argued (Huff, Rattner, and Sagarin 1996). In fact, the very small percentage of guilty pleas holds even for exonerations resulting from events that are likely to be independent of defendant behavior – that is, when either the real offender or the alleged murder victim were found. Even in this subset of exonerations, one finds only 6.4% of guilty pleas (12 of 186). Moreover, this percentage is not significantly different from the remaining 7 of 130 exoneration cases in the database (5.8%) that followed a guilty plea ($\chi^2 = .15, p = .69$).

The dramatic and systematic difference in the guilty plea rates of exoneration cases versus the broader defendant population was also echoed in the behavior of the 38 Tulia scandal defendants described above. Given their special circumstances, these defendants exhibited a relatively high 71% guilty plea rate as a group. Nevertheless, even this understandably high rate is significantly lower than the approximately 95% guilty plea rate that obtained at the time for small Texas counties comparable to Tulia's Swisher county.  

Thus, both our analysis of exoneration statistics and the anecdotal evidence suggest innocent defendants are more likely to reject plea offers. For the plea bargaining debate, however, this suggestive evidence is incomplete. There is no telling how common stories like the Jarrett or Hayes anecdotes are, and even the striking exonerations statistics and the Tulia defendants' behavior may be explained on rational grounds, without reference to fairness considerations. For example, innocent defendants may possess private information about their better chances at trial (Froeb 1993), information they are unable convincingly to convey to prosecutors. If this were the case, innocents would rationally reject some offers their guilty counterparts would accept. In fact, such a finding would not be surprising; after all, the basic function of the criminal trial is to distinguish between the innocent and the guilty (Grossman and Katz 1983, Scott and Stuntz 1992a). Furthermore, beyond the rational account, innocents may hold optimistically biased predictions of their conviction odds and likely sanction at trial (e.g., Bibas 2004). There is also experimental evidence these defendants are more optimistic about their conviction probability than their guilty counterparts, even when faced with identical evidence (Bordens 1984; Gregory, Mowen, and Linder 1978).

---

10 We used the Bureau of Justice Statistics (2001) to compile the rate of guilty pleas in felony cases in rural counties with fewer than 50,000 inhabitants in Texas, which resemble Tulia's Swisher County.

11 If this information is not private, guilty pleas should be as likely in weak cases as in strong cases, since prosecutors would adjust their plea concessions and obtain guilty pleas regardless of the probability of trial convictions.

12 For a discussion of overoptimistic predictions in plea bargaining see infra Part 2.2.2.
All in all, rational probability estimates and over-optimism may or may not join with considerations of fairness to cause rejections of plea offers. For the purpose of legal analysis and policy formulation, however, it would be beneficial to disentangle the impact of these different variables on plea behavior. For instance, law and economics scholars who champion defendants' freedom to bargain may find rejections based on rational probability estimates desirable and at the same time wish to debias defendants who are overoptimistic (Scott and Stuntz 1992a, Bibas 2004). Such scholars may similarly consider acceptable plea rejections based on a defendants' preference for substantive fairness, but object to rejections following erroneous judgments of the unfairness of some plea offers.

For the task of clarifying the role of different variables in plea rejections, controlled experimental tests are better suited than field studies. The following section thus examines the important but limited extant experimental findings on the role of substantive fairness considerations in plea bargaining.

### 1.2. The Early Plea Bargaining Experiments

Behavioral decision making scholars would not be surprised by the suggestive real world evidence on the role of fairness concerns in plea bargain decision making. In fact, the broader behavioral literature provides ample evidence of the impact of fairness perceptions and related concerns on behavior in bargaining and allocations outside the present domain (e.g., Bazerman, Loewenstein, and White 1992; Camerer 2004; Kahneman, Knetsch, and Thaler 1986; Roth 1995). Two early experimental studies, however, specifically examined plea bargain decision making in designs that implicate the impact of substantive fairness concerns on WTAP.

#### 1.2.1. Gregory et al.'s (1978) Study

In two experiments, Gregory et al. (1978) found substantive fairness impacted WTAP. The first study asked male students to imagine they were either innocent or guilty of having committed armed robbery. They received highly detailed information on the circumstances that led to their arrest for committing the crime, the charges against them, their punishment if convicted, and details of a plea bargain they were offered. The first three variables – namely, culpability, number of charges, and prison sentence if convicted – were manipulated between subjects, while plea offers were identical for all participants. On the main measure of interest for our purposes, this study found innocent participants significantly more likely than guilty ones to reject the plea offer.\(^{13}\)

A second, follow-up, experiment used a confederate to place students in conditions of actual guilt or innocence about having prior information on answers to a difficult test. This more realistic experiment corroborated the results of the hypothetical judgments and decisions made by participants in the first experiment. It found innocents dramatically less likely than guilty participants to accept a plea-like compromise offer in lieu of facing judgment by an ethics committee with potential higher penalties. Together,

\(^{13}\) Interestingly, the study also found overall main effects for the other independent variables (i.e. number of charges and severity of punishment) and an interaction between them. These effects, however, appeared when analyzing guilty but not innocent participants. This suggestive evidence further corroborates the dominance of substantive fairness in innocents' decision making.
these two controlled experiments complement the field evidence and indicate that, at least in the plea bargain domain, hypothetical studies using students as participants and require "as if" behavior may well possess external validity.

Despite their importance, however, Gregory et al.'s (1978) findings are subject to some limitations. Specifically, those innocent participants in the first experiment, who were reluctant to accept plea offers, also believed they were less likely to be convicted and that their (hypothetical) defense attorney and judge would be more likely to think them innocent. Such participants would have found plea offers less attractive even in the absence of any substantive fairness concerns. The same holds for participants in the second, more realistic, experiment, whose behavior might have reflected this confound between the effects of fairness preferences and probability judgments. In fact, the guilty among the second experiment participants probably did face a higher probability of conviction than the innocent ones. The former knew a potential witness (the experimental confederate) might incriminate them, while the latter had no such concerns. Overall, therefore, Gregory's (1978) studies leave some uncertainty as to the independent impact of substantive fairness on WTAP.  

1.2.2. Bordens' (1984) Study

Bordens (1984) extended Gregory et al.'s (1978) findings to gain better understanding of defendants' plea bargain decision making. Using a large factorial design he independently manipulated culpability, sentence offered in the plea bargain ("Offer"), a defense attorney's estimate of conviction likelihood ("Conviction Probability"), and the sentence defendant would receive if convicted at trial ("Sentence").

Participants were provided with a description of the crime with which they were charged (negligent homicide by automobile) and the evidence against them (three eyewitnesses). They were told whether they were innocent or guilty, and given a scale measuring their initial likelihood of conviction estimates, which they were then explicitly told to disregard. Thereafter they were given information on their Offer, Conviction Probability, and Sentence. Bordens (1984) used four levels of Offer (probation, 6 months in prison, one year in prison, and three years in prison), three levels of Conviction Probability (10%, 50%, and 90%), and two levels of Sentence (one year more or five years more than Offer).

The results showed Bordens' (1984) innocent defendants significantly more likely than their guilty counterparts to reject plea offers. Before being provided with the defense attorney's estimates, moreover, innocents believed they were significantly less likely to be convicted at trial than did guilty participants. Unsurprisingly, the independent

---

14 While focusing on defendants' preferences regarding procedural aspects of plea bargaining, Houlden (1980) sought to address this confound in Gregory et al.'s (1978). She also used as participants both students and inmates at a Florida country detention center. This experiment included four car accident scenarios, and respondents were asked to imagine that they were the driver and had been charged with first-degree murder. The scenarios manipulated culpability and the strength of evidence against the driver. Participants' preferences for and perceptions of different plea bargaining procedures were measured. They also completed manipulation checks of their perceptions of culpability and strength of the evidence against them. However, Houlden (1980) neither compared the overall responses of guilty and innocent participants nor provided sufficient information (i.e. number of participants in cell and standard deviations) to allow one to make such comparison based on the reported data.
variables other than culpability all had significant main effects on WTAP as well, with willingness to plea increasing with Conviction Probability and Sentence and decreasing with severity of Offer.\footnote{Similar findings were obtained using an additional dependant variable of "firmness of decision" that was used to complement the binary accept/reject choice regarding the plea offer.}

These findings corroborate the field evidence and Gerogry et al.'s (1978) results, but their usefulness for the present day plea bargaining debate is still limited in some respects. Bordens' (1984) design overcame the earlier confound of preference and probability judgments by directly manipulating conviction probabilities. At the same time, however, the independent manipulation of Conviction Probability and Sentence meant this design could not control for the impact of expected sanction – the joint product of Conviction Probability and Sentence – on WTAP.\footnote{One level of the Offer variable (probation) was also on a different scale from that used in the other Offer levels and the Sentence variable alike and thus incompatible with an expected value analysis.}

Consequently, offers were higher than expected sentences in some experimental conditions and lower in others. In the 10% Conviction Probability conditions, for example, the Offer was almost always higher than the expected sentence.\footnote{Except for those cells where 6-months (or 0.5 years) offers had to be traded off against a 10% probability of 5.5 years (expected value of 0.55 years).} In the 90% Conviction Probability conditions, on the other hand, the Offer was always lower, and often much lower, than the expected sentence.\footnote{The same holds for most 50% Conviction Probability cells.} Given this state of affairs, it is difficult to draw clear conclusions from the reported results regarding the main effects of the independent variables other than culpability or from the interactions among the different variables. The main effect found for Conviction Probability, for instance, reveals little about participants' risk preferences; after all, Offers would have been dramatically more attractive in the high – as compared to the low - Conviction Probability conditions even for perfectly risk neutral defendants.\footnote{Bordens (1984) also does not provide sufficient cell data, rendering further reanalysis impossible.}

In sum, Bordens' (1984) findings provide important evidence for the independent role of substantive fairness on WTAP. They also suggest that defendants may consider both Conviction Probability and Sanction when choosing between Offer and trial, but do not control for the impact of the expected sanction on WTAP. These results therefore leave open the questions of defendants' risk preferences in plea bargain decision making and, importantly, of the interaction between these preferences and the substantive fairness of plea offers. We therefore turn, in Part 2, experimentally to examine the impact of substantive fairness concerns on defendants risk preferences and WTAP, which has important implications for present day plea bargaining scholarship.

2. REEXAMINING SUBSTANTIVE FAIRANESS IN PLEA ACCEPTANCE

Early experimental studies showed that the results of hypothetical questionnaires largely parallel those of complex simulation studies (Gregory et al. 1978) in the context of plea acceptance, and that the responses of students mostly resemble those of real-world inmates (Houlden 1980). We have also shown that, together with varied field evidence,
the early studies indicate substantive fairness may affect WTAP, but remain somewhat inconclusive. Against this backdrop, the present experimental tests use hypothetical questionnaires with college and law student participants. We complement the extant evidence with simple scenarios that present our student participants with situation they plausibly could face in the course of their daily lives. These scenarios control for those variables that are most important for present-day legal analysis, providing further insight into the role of fairness in pleas. Significantly, moreover, the framework developed here highlights the importance of additional decision making phenomena for plea bargaining scholarship.

2.1. Study 2: The Impact of Substantive Fairness

Study 2 tested the hypothesis that substantive fairness impacts defendants' willingness to accept plea bargain offers, controlling for the expected sanction at trial. This control, in turn, allowed us to isolate the impact of preferences for substantive fairness, avoiding potential confounds with both subjective judgments of and objective differences in expected trial outcomes. The control for expected sanction also permitted us directly to contrast our findings with prevalent models of rational defendants' behavior. Furthermore, Study 2, like all of the remaining studies reported here, focused on WTAP as a dependent measure that is directly relevant for the legal plea bargaining debate. In line with the earlier plea bargaining experiments (Bordens 1984; Gregory et al. 1978) and to prevent problems of demand characteristics we avoided introspective reports of perceived fairness and manipulated culpability between subjects only. We predicted that culpable defendants will exhibit a greater WTAP than innocents, holding fixed all other plea variables (i.e. Offer, Conviction Probability, and Sentence).

A total of 30 undergraduates from the University of Michigan (17 females and 13 males) were recruited at the library to volunteer for a short questionnaire survey. In a between-subjects design, participants read about being charged for an academic violation that they either committed or did not commit: “Imagine you were accused of cheating on an exam (and personally you know you were [not] cheating).” Then all participants read, “If your case goes before the Standards and Ethics Committee at UM, given your best ability to present your case, you see your odds, at best, being the following: 60% chance of being suspended for 1 year from UM or 40% chance of exoneration. However, an alternative is to appeal your case before it goes to the Standards and Ethics Committee but the consequences are that you would receive a ‘FAIL’ for the class but would not be suspended from UM.”

At this point, all participants answered the following dependent variable: “Strategically speaking, what would you choose? (please check).” The two options were “Appeal the case and receive a ‘Fail’ for the class – OR – Allow the case to go the Standards and Ethics Committee.”

The results were consistent with the prediction ($\chi^2=6.65$, $p<.05$). Among those who were told they did cheat, 67% accepted the plea arrangement of appealing the case and receiving the Fail for the class. In contrast, only 20% of those who did not cheat accepted this arrangement; the overwhelming majority in this condition chose to have the case go before the Standards and Ethics Committee, despite the odds of suspension.
These findings reflect the basic impact of substantive fairness in a between-subjects setting, controlling for the outcomes of plea acceptance and trial before the Committee, respectively. The results are unlikely to have been driven by a disparity in predictions of conviction probability between the guilty and innocent conditions: The scenario that informed participants of their subjective probability estimates left little room for alternative predictions. Moreover, in accordance with the approach we outlined above, the study purposefully did not provide participants with rich information about the event and surrounding circumstances (cf. Babcock and Loewenstein 1997; Loewenstein and Moore 2004), which might have contributed to biased predictions of trial outcomes in the two earlier studies (Bordens 1984; Gregory et al. 1978). We therefore attribute the significant difference between the two conditions to the experimental manipulation of substantive fairness.

2.2. Study 3: Biased Assessments of Substantive Fairness

Defendants whose WTAP depends on their substantive fairness perceptions must first determine whether they are culpable. Frequently, this determination is easy. A defendant who consciously engaged in illegal activity is likely to have few doubts about his culpability. Nevertheless, real defendants are sometimes uncertain of their culpability (e.g., Bibas 2004). They may be uncertain, for instance, of the facts with which they are charged. A driver may not be sure whether he exceeded the posted speed limit when stopped by the police, since he was not paying attention to the speedometer at that time. Similarly, a person who was drunk at a party and is charged with assaulting another guest may be uncertain as to whether he indeed assaulted that other person or whether he had done so in self defense. In fact, the drunken defendant may even be uncertain whether the assault had ever happened.

Even a rational uncertain defendant with a taste for substantive fairness might still consider his uncertainty irrelevant to the plea decision upon lacking necessary information about his culpability. A truly uncertain defendant may also estimate his likelihood of culpability at 50%, in the absence of further information, consequently finding his fairness concerns less acute. Alternatively, the uncertain defendant may subjectively judge his culpability based on the available information and express a preference for substantive fairness that reflects that judged culpability. Under all of these scenarios, however, rational uncertain defendants would not exhibit an aversion to plea bargaining resembling that of innocents. At most, they would show some intermediate level of plea rejections, reflecting their probabilistic judgment that a plea is substantively unfair.

In contrast with the rational prediction, the behavioral decision making literature suggests that uncertain defendants may exhibit an egocentric bias in their culpability judgments. Egocentrically biased defendants disproportionately will deem themselves innocent under uncertainty and will therefore reject pleas offers as if they were innocent. Such biased defendants, would still face higher average sanctions than guilty defendants.

---

20 Note the Bordens (1984) found biased probability judgments before attempting to control for the bias by providing participants with their defense attorney estimates regarding trial.

21 Except possibly where the objective evidence strongly points out to their innocence, a case in which a prosecution is unlikely.
– much like the "cost of innocence" borne by their innocent counterparts – but without a rational justification for their worse outcomes.

Moreover, evidence of an egocentric bias in WTAP would also make an important contribution to the legal debate over the innocence problem, where scholars have argued that uncertain defendants are more likely to plead guilty because they lack valuable information (Bibas 2004; Scott and Stuntz 1992a). Evidence of egocentrically biased plea rejections by uncertain defendants – who have no private information – would also reinforce our Study 2 finding that such information is not necessary to generate plea rejections.

We thus provide a brief review of the evidence on egocentric judgments in bargaining generally and plea bargaining specifically, followed by Study 3, which compared uncertain to certain defendants.

2.2.1. Egocentric Judgments of Fairness in Bargaining

The extensive literature on egocentric judgments (e.g., Allison, Messick, and Goethals 1989; Dunning, Heath, and Suls 2004; Dunning, Meyerowitz, and Holzberg 1989; Epley and Dunning 2000; Griffin, Dunning, and Ross 1990) suggests that uncertain defendants will tend to view themselves in a positive light. Such defendants will behave as if innocent and therefore reject plea offers that unbiased defendants who are uncertain of their guilt would accept.

The experimental evidence on egocentrically biased judgments of fairness in civil litigation and settlement (e.g., Babcock and Loewenstein 1997; Babcock, Loewenstein, Issacharoff, and Camerer 1995; Loewenstein, Camerer, Issacharoff, and Babcock 1993; Messick and Sentis 1979) further supports this hypothesis. For instance, experimental participants were shown to interpret civil trial evidence egocentrically depending on their designation as plaintiffs or defendants, and real-world negotiators in labor disputes identify fair targets for comparison in egocentric ways. These parties' biased judgments of fairness have also been linked to a diminished likelihood of settlement between them (Babcock and Loewenstein 1997).

2.2.2. Evidence of Egocentric Predictions in Plea Bargaining

The extant experimental studies of plea bargaining did not examine whether evaluations of plea offer fairness are egocentrically biased, but indicate innocents may exhibit biased predictions more generally. Gregory et al.'s (1978) innocent participants evaluated their odds of acquittal at trial as significantly higher than did guilty participants. The former also exhibited a dramatically stronger belief than the latter that their defense lawyer believes their version of the events. Since the guilty and innocent experimental conditions only differed in the participants' private knowledge of their respective innocence or guilt, it is possible that the aversion of innocents to a substantively unfair outcome contributed to their optimistic view of their ability to convince their lawyer and judge of their innocence. However, these defendants may have rationally believed they would be better able to convince others of their true innocence.

22 We find the oft-used term "self-serving bias" misleading. Egocentric biases may protect self-perceptions, and are thus narrowly "self-serving," but their consequences may well harm the biased decision maker.
A similar pattern was found by Bordens (1984), with innocents more optimistic about their trial odds. In this study innocents also had some limited rational basis for believing their trial odds better. They were given detailed scenarios regarding the events preceding their hypothetical arrest and were told there were 3 witnesses to the alleged crime. Participants' biased estimates also were made before an experimental control instructed them "to disregard their own estimates of the likelihood of conviction."

However, the illusion of transparency (Gilovich, Savitsky, and Medvec 1998) may better account for the difference between the predictions of culpable and innocent participants in both Gregory at al. (1978) and Bordens (1984) (before the experimental control). This illusion refers to decision makers' tendency to overestimate the extent to which others can discern their internal states. People often mistakenly believe their internal states are more transparent than they really are. Gilovich et al. (1998) show, for example, that liars overestimate the ability of others to detect their lies, a finding that may well explain the systematic difference between predictions of trial outcomes by guilty and innocent defendants. The guilty inevitably must lie to claim innocence, and may overestimate the ability of prosecutors, judges, and juries to recognize their lies. At the same time, the illusion of transparency should exert the opposite effect on innocents, leading them to overestimate the ability of the legal system to identify their innocence.

In sum, while there is much data on egocentric biases generally and in litigation and bargaining more specifically, the evidence in the plea bargaining domain is tentative at best. No previous study, moreover, tested the impact of egocentric assessments on defendants' judgments of culpability and fairness or linked such biased judgments to plea acceptance – the focal measure in the present set of studies. Such a test, however, could show the impact of egocentric biases in pleas while ruling out potential confounds, either with rational judgments of conviction probabilities or with the illusion of transparency.

2.2.3. Our Study: Uncertain Offenders

Study 3 tested the hypothesis that defendants will exhibit egocentric assessments of culpability in the plea bargain setting. We expected participants who are uncertain of their culpability to behave as if they were innocent – namely, as if their plea offers were substantively unfair. For the reasons explained in Study 2, we focused on the WTAP measure and refrained from asking for participants' introspections regarding the fairness of offers. Importantly, Study 3 also placed the expected outcomes of trial and plea, respectively, on the same sanction scale, allowing a comparison of participants' risk attitudes not only across conditions, but also to theoretical predictions of rational defendant behavior.

A total of 59 undergraduates from the University of Michigan (31 female and 28 male) were recruited at the library to volunteer for a short questionnaire survey. In a between-subjects design, participants read about accidentally hitting a pedestrian and were assigned to one of three conditions that varied substantive fairness:

After turning a sharp curve, you saw a woman slowly crossing the road. You tried to break as quickly as possible, but your car hit the woman, who later died from her injury. Personally, you (know that you clearly / are unsure whether or not you / are certain that you did not) exceeded the speed limit.
All participants then read,

You now face criminal charges, and the outcome of your trial depends on whether the prosecution proves that you exceeded the speed limit. Your highly experienced lawyer reviewed the evidence and consulted with other experts and tells you that your odds are even: A 50% chance of being convicted and getting a mandatory 4-year suspension of your driver’s license OR A 50% chance of not being convicted and walking free.

Participants continued,

Your lawyer explains that only he and the prosecutor will meet just before the trial and that the prosecutor will offer a plea bargain, where he will drop the current charges if you will plead guilty to a lesser offense that carries a significantly lighter sentence. However, you must now tell your lawyer the maximum license suspension you are willing to accept as part of the plea bargain, to avoid the 50% risk of getting the mandatory 4-year suspension at trial.

At this point, participants responded to the dependent variable, “What is the maximum period of license suspension you would be willing to accept as part of the plea bargain, to avoid the 50% risk of getting the mandatory 4-year suspension at trial? (please check)” They then indicated the maximum sanction they were willing to accept (“ACCEPT”) on a scale from 0 to 48 months with 6-month intervals and with the midpoint of the scale representing 24 months.

As predicted, participants in the guilty condition were willing to accept a higher maximum sanction ($M=3.95$, $SD=2.06$) than those in both the not guilty ($M=2.68$, $SD=2.08$) and uncertain conditions ($M=2.64$, $SD=2.37$). We conducted a linear regression to test our prediction with the contrast weights: 2 (guilty), -1 (not guilty), -1 (uncertain). The results were significant ($B=-1.29$, $β=-.28$, $p<.05$), while controlling for the orthogonal contrast (0, 1, -1: $p=.96$).

The results implicate participants' egocentric assessments in the uncertain condition. These participants' diminished ACCEPT ratings closely resembled those of participants in the not guilty condition, and were significantly different from those of participants in the guilty condition. Apparently, when faced with uncertainty regarding their guilt, participants tend to behave as if they were not guilty. This pattern cannot be explained by rational judgments of conviction probabilities or by the illusion of transparency, since conviction probabilities were identical across conditions. Therefore, it implicates participants' egocentric self perceptions. The tendency of uncertain defendants in Study 3 towards plea rejection is in clear contrast with models of rational defendant behavior, however. Even uncertain rational defendants holding a preference for substantive fairness would not mimic so perfectly the behavior of innocents.

In addition to revealing the potential impact of egocentric biases on WTAP, our ACCEPT measure also suggests that guilty participants exhibit risk attitudes that differ

---

23 Even absent a control for conviction probabilities, the illusion of transparency probably should not have led uncertain defendants to predict that others will judge them innocent.
from those of innocent and uncertain ones. In the scenario we tested, the expected license suspension sanction at trial was 2 years (0.5 probability of a 4-year mandatory suspension), indicated as 24 months or 4 on the ACCEPT 8-point scale. Risk averse participants would therefore have exhibited an ACCEPT > 4, risk neutral ones an ACCEPT = 4, and an ACCEPT < 4 would have reflected risk seeking. Interestingly, the choices of participants in the guilty condition resembled those of a risk neutral decision maker (t(19) = -.1, p=.92), while participants in the not guilty and uncertain conditions, on the other hand, exhibited significant risk seeking (t(38)= -3.9, p<.001).

While intriguing, however, participants' absolute risk attitudes in Study 3 may have been an artifact of the specific design. The results of this study did replicate the findings of Study 2 and corroborated the earlier evidence on diminished WTAP, which means greater risk taking on the part of innocents. However, this increased risk taking may reflect innocents' lower degree of risk aversion rather than any risk seeking behavior on their part. Study 4 therefore sought to probe the impact of substantive fairness concerns on defendants' risk attitudes for different levels of conviction probability.

2.3. Fairness and Risk Attitudes in Plea Acceptance

2.3.1. Fairness and Reference Standards

Studies 1 and 2 showed that defendants are averse to accepting plea offers that violate their standards of substantive fairness. Study 3 further indicated that such standards may serve as reference points that make innocent defendants more risk seeking than their guilty counterparts. Specifically, we suggest that defendants' aversion to substantively unfair outcomes leads them to view outcomes as positive or negative depending on their perceived fairness: Substantively unfair outcomes are viewed as negative, while fair outcomes are either neutral or positive. According to this account, guilty defendants would tend to view conviction as fair and a discounted plea offer as a positive outcome vis-à-vis this neutral reference point. Innocents, on the other hand, will tend to view acquittal as fair, while perceiving conviction and sanction as unfair, negative, outcomes compared to that neutral reference point. This hypothesis is consistent with Kahneman et al. (1986), who show that decision makers judge the fairness of outcomes based on their coding as gains and losses relative to a psychologically neutral reference point.

Kahneman et al. explain that judgments of fairness are analogous to individuals' preferences, which are evaluated as either gains or losses relative to a reference point and are susceptible to framing effects (e.g., Kahneman and Tversky 1979; Kahneman and Tversky 1984). They show, for example, how participants in household surveys judge differently the fairness of two comparable salary reductions: A majority of participants found a 7% wage cut unfair, but a large majority also found a 5% nominal salary raise in the face of a 12% inflation to be fair (Kahneman et al. 1986). These authors attribute this and related findings to the differential psychological effects of gains and losses relative to

---

24 There is evidence that defendants, arrestees, and members of the general public are hyperbolic discounters (i.e. risk-seeking) when comparing alternative sanctions over time – such as a one versus four years of imprisonment (Spelman 1995). These findings, however, are of limited relevance to the present question, since they do not distinguish between fair and unfair sanctions, which the variable we focus on here.
a reference point, which make objectively comparable outcomes appear more unfair when coded as a loss than when coded as the elimination of a gain.

We thus argue that the analogy Kahneman et al. (1986) draw between the psychology of fairness judgments and the psychology of individual choice has important implications in the plea bargain domain and elsewhere. Specifically, reference points that shape fairness judgments, may also impact fairness-driven choices. In the domain of individual choice, there is much evidence that decision makers exhibit risk aversion for gains but risk seeking for losses (Kahneman and Tversky 1979; Tversky and Kahneman 1991). If these phenomena were to apply to fairness-driven choices, we would expect fairness reference points to generate different risk attitudes for "positive" and "negative" outcomes in plea bargains, as indicated by Study 3: Guilty defendants, who generally view plea offers as gains, should exhibit risk aversion. Innocents, on the other hand, view the same offers as losses and therefore exhibit risk seeking choices in plea bargaining.

Importantly, moreover, criminal defendants must trade off (fair or unfair) plea offers against expected trial outcomes, which are probabilistic. Their choice between the plea and trial alternatives is therefore made under uncertainty. However, more recent work by Tversky and Kahneman (1992) on individual choice under uncertainty shows that risk attitudes depend not only on the coding of outcomes vis-à-vis the neutral reference point, but also on the probability of these outcomes: Medium to high probability gains and small probability losses generate risk aversion, while medium to high probability losses and small probability gains lead to risk seeking choices. When translated into the present domain, these findings imply that defendants' risk attitudes depend not only on plea offer fairness but also on the probability of acquittal at trial, which is clearly the outcome preferred by all defendants.

To wit, guilty defendants, who view plea offers as gains, should exhibit risk aversion for medium to high, but risk seeking for very low, acquittal probabilities. On the other hand, innocents should be risk seeking when acquittal probabilities are medium to high, but risk averse when these probabilities are very low. Notably, the findings of Study 3, where innocents (and uncertain defendants) exhibited risk seeking, while guilty participants were risk averse, under a uniform medium 50% acquittal probability, also provides some support for this fourfold hypothesis.25

2.3.2. Study 4: Substantive Fairness and Risk Attitudes

Study 4 examined participants' risk attitudes for substantively fair and unfair plea offers, using a within-subjects design to underscore the strength of individual preferences (Camerer 1995). As before, we hypothesized that guilty participants will view the plea offer as substantively fair, while innocents will view it as unfair. The former were thus predicted to view the plea as either neutral or a certain gain, while to the latter it was predicted appear a certain loss. Therefore, similarly to advanced prospect theory (Tversky and Kahneman 1992), we expected guilty participants to be risk neutral or risk averse for medium to high, but risk seeking for low, probabilities of acquittal at trial. At the same

25 Bordens' (1984) early finding of a marginally significant interaction between culpability and estimated probability of conviction is also suggestive. That study, however, neither systematically varied probability and trial sanction versus plea offer nor controlled for expected sanction.
time, innocents were expected to exhibit risk seeking for medium to high, but risk aversion for low, acquittal probabilities.

In addition, Study 4 controlled for the effect of the expected sanction at trial – against which the participants had to trade the plea offer – by holding the plea offer equal to the expected sanction. This design was also chosen to make absolute risk attitudes transparent: Risk averse defendants would always accept the plea offer, to avoid the risk of facing the higher sanction in case of conviction at trial. Risk seeking participants would always reject the plea offer, which offers them no "discount" from the expected sanction at trial. And risk neutral defendants would be indifferent to the choice between plea and a trial with an identical expected sanction value.

A total of 64 undergraduates from the University of Michigan were recruited to participate in an online questionnaire survey. Approximately 200 e-mail addresses were randomly targeted from the student directory; the response rate was 31 percent.

In a within-subjects design, participants read a variant of the car accident scenario used in Study 3:

After driving around a sharp corner, you hit a person who died as a result. You now face criminal charges, and personally you know you (DID exceed / did NOT exceed) the speed limit. The outcome of your trial depends on whether the prosecution proves that you exceeded the speed limit. If they succeed, you will be convicted and will get a mandatory 5-year (60 months) suspension of your driver's license. If they fail, you walk free.

Participants were then asked to decide whether to accept a plea bargain in 5 different circumstances: “Now, just before the trial, the prosecutor offers you a plea bargain: he will drop the current charges if you will plead guilty to a lesser offense that carries a significantly lighter sentence. If the terms were the following...”

At this point, participants read their 5 plea offers, in which they were asked to choose between going to trial, with a given sanction and probability of conviction, and accepting the plea offer. In each case, the sanction included in the plea bargain was set to equal the expected sanction at trial. Participants then chose between a 5% probability of getting the full 5-year suspension and a 3-months plea-bargain; a 30% probability of conviction and an 18-months offer; a 50% probability and a 30-months offer; a 70% probability and a 42-months offer; and, finally, between a 95% probability and a 57-months offer. The order of the 5 offers was randomized, and the question “...which would you choose?” preceded each offer.

To test these correlated binary response data, we conducted a GEE analysis (Liang and Zeger 1986). The results were consistent with the prediction. There was a main effect for GUILT ($\chi^2=7.60, p<.01$) – that is, participants were significantly more likely to reject a substantively unfair offer than a substantively fair one, over all levels of

---

26 Because the binary logistic regression assumes that the responses (Trial / Plea) are independent, we could not use that procedure in the present within-subjects design. The appropriate test for these correlated binary responses is the Generalized Estimating Equations (GEE) Analysis (Liang and Zeger 1986). We conducted our GEE analyses using the REPEATED statement in the GENMOD procedure in SAS.
probability. There also was a significant main effect for PROBABILITY ($\chi^2=16.5$, $p<.01$), indicating differential effects for different levels of probability of conviction.

As Figure 1 clearly illustrates, participants in the innocent condition tended to exhibit significant risk seeking and reject the plea offer across the probability range, with plea acceptance rates of only 17%, 7%, 20%, 43%, and 50% for probability of acquittal levels of 95%, 70%, 50%, 30%, and 5%, respectively. Guilty condition participants, on the other hand, were significantly more risk averse, with plea acceptance rates of 47%, 38%, 56%, 53%, and 41% for the same probability of acquittal levels of 95%, 70%, 50%, 30%, and 5%, respectively.

![Figure 1](image.png)

Figure 1. Car accident: percent accepting plea offer by probability of conviction

Hence, innocent participants were never risk averse on average, and only exhibited risk neutrality for the 5% probability of acquittal. Guilty participants, however, never exhibited strong risk seeking, and were approximately risk neutral on average. Our hypothesis regarding the joint effects of substantive fairness reference standards and acquittal probabilities was thus generally confirmed.

Moreover, there also was a significant GUILT x PROBABILITY interaction ($\chi^2=12.3$, $p<.05$), suggesting that changes in probability levels differently impacted participants in the guilty and innocent conditions. A visual examination of Figure 1 clearly shows that acquittal probability generated parallel trends for the guilty and innocent conditions in the high to medium (95%-50%) probability levels. However, as acquittal became less likely (30%) then extremely unlikely (5%), the behavioral trends of participants in the two conditions diverged. Innocent condition participants – who were generally risk seeking - became more risk averse, while guilty condition participants –

---

27 We present and analyze acquittal probabilities although the questions used the (complementary) conviction probabilities, since the former focus on the variable of interest in prospect theory terms, while we thought the latter phrasing more fitting for a rational tradeoff between the probable sanction on trial and the sure sanction demanded by the plea offer.
who were generally more risk averse – became increasingly risk seeking. This result generally supports our hypothesis regarding the applicability of the fourfold risk attitude pattern of advanced prospect theory (Tversky and Kahneman 1992) to fairness-driven choices in plea bargaining.

This predicted pattern, however, better explains the respective behavioral trends for guilty and innocent participants than the absolute risk attitudes these two groups exhibited for low acquittal probabilities. Innocents in these conditions exhibited an increased preference for plea over trial, revealing a diminished risk seeking but not reaching absolute risk aversion. Guilty participants similarly showed risk seeking, but only at a modest level. However, the muted level of absolute risk aversion and risk seeking in this study may have resulted from specific experimental design feature: To fit Study 4 with both the extant plea bargaining experiments and the theoretical literature on rational defendant behavior, participants were presented with conviction, rather than with acquittal probabilities. The probability format was thus controlled for, but not manipulated, in our experiments. Decision making research in other domains, however, shows how the language used in question framing generally and probability format specifically can impact choice (e.g., McNeil, Pauker, Sox, and Tversky 1982). It therefore, possible that a design highlighting "acquittal" instead of "conviction" would have made absolute risk attitudes more extreme than those found here,\textsuperscript{28}

Study 4 therefore shows that substantive fairness standards may provide reference points for defendants' plea behavior. Guilty defendants, who view the reduced penalty in the plea offer as a gain vis-à-vis the fair outcome, tend to be more risk averse. Innocents, on the other hand, must trade two unfair outcomes against one another. When choosing between the sure loss in the plea and the risk of the higher loss at trial, these defendants tend to be more risk seeking. These risk attitudes reverse their course for low to very low acquittal probabilities, largely – although imperfectly – resembling the fourfold pattern of risk attitudes predicted by advanced prospect theory.

In sum, Studies 2-4 adduce important additional evidence for the independent impact of substantive fairness on WTAP. The power of reference standards was further explored in Studies 3-4: Study 3 showed participants who are unsure of their culpability exhibit an egocentric bias, manifested in an aversion to plea acceptance equaling that of innocents. Study 4 then revealed that standards of substantive fairness may shape risk attitudes across the probability range. Innocent and guilty participants, respectively, view plea offers within loss and gain frames following their different perceptions of what constitutes a neutral, "fair," reference point. These differing reference points, in turn, generate different risk attitudes in the two defendant types across the probability range. Together, Studies 2-4 also provide, first, meaningful experimental tests for pervasive – though inaccurate – assumptions in present-day's plea bargaining discourse regarding defendants' plea behavior; and, second, an important link between the experimental analysis of plea bargaining and the broader behavioral decision making literature.

\textsuperscript{28} One might also speculate that if incarceration were used as the sanction in this design instead of license suspension, we might have observed increased risk seeking overall given the likely more extreme aversion of normative participants to jail.
3. THE COMPARATIVE EVALUATION OF PLEA OFFERS

3.1. Fairness and Comparative Evaluation

Comparative evaluation – that is, the evaluation of an offer vis-à-vis offers made in similar cases - provides defendants with another measure of the fairness of the proposed deal. This evaluation naturally may lead defendants to conclude their offer is superior, inferior, or comparable to other offers. Subjectively, inferior offers likely would be deemed unfair, while both superior and comparable offers probably would not raise a similar concern. However, superior offers may also appear more attractive, compared to those merely "fair," neutral, comparable offers. Therefore, to capture this distinction – between attractive and merely fair offers – we refer to the "comparative evaluation," instead of "comparative fairness," of plea offers.

Models of rational defendant behavior consider the comparative evaluation – much like the substantive fairness – of offers irrelevant for plea choice, since it changes neither the terms of the offer nor the rational defendant's risk preferences. Real defendants, however, may determine the attractiveness of their offers also by comparing their terms to other offers, finding comparatively inferior offers less attractive and comparatively superior ones more attractive than they would otherwise be. Comparatively inferior offers in allocation decisions may violate common notions of fairness as equity or equality (e.g., Deutsch 1975; Rawls 1971, pp. 235-237). They are often described or experienced as "unfair" by their recipients.

We have already seen, moreover, that decision makers view allocations below those "reference transactions" that are perceived as relevant comparisons for the present transaction's fairness as unfair (Kahneman et al. 1986). Similarly, the extensive ultimatum game literature shows allocation recipients' ("Responders") willingly to sacrifice monetary payoffs when offered a comparatively unfair share by their counterparts ("Proposers") (e.g., Camerer, 2004; Güth, Schmittberger, and Schwarze 1982; Roth 1995). The independent role of fairness-related concerns – as opposed to strategic considerations – in this allocation setting is further highlighted by those variants of the ultimatum game that show Responders to reject comparatively inferior allocations even where their rejections have no impact on Proposers' payoff (e.g., Bolton and Zwick 1995; Fong and Bolton 1997). Accordingly, other studies reveal the role played by Responders' negative emotional reactions to comparatively unfair allocations (e.g., Falk, Fehr, and Fischbacher 2003; Kirschsteiger 1994; Pillutla and Murnighan 1996).

Comparative evaluation may be relevant, however, where it provides defendants with information regarding their bargaining power vis-à-vis the prosecution, although this strategic consideration is beyond the scope of the present analysis. Our comparative evaluation studies therefore examine only one-shot, non-negotiable, offers, where strategic considerations should be irrelevant.

More recent experimental evidence suggests that other negative emotions beyond pure fairness concerns contribute to recipients' aversion to comparatively disadvantageous outcomes and their preference for comparatively advantageous ones (e.g. Pillutla and Murnighan 1996; Pillutla and Murnighan 2003). This is another reason for using here the more general "comparative evaluation" terminology, which encompasses a variety of motivations beyond fairness considerations alone.
Furthermore, while ultimatum game Responders reject comparatively unfair divisions made by their peers, choice studies reveal that decision makers are averse to disadvantageously unequal allocations per se, even when making the allocation decision themselves (e.g., Loewenstein, Thompson, and Bazerman 1989). Between subjects, separate evaluation, participants in Lowenstein et al.’s (1989) study, for example, rated different options that provided them and another person with certain monetary sums. One group of participants rated the outcome of $500 for oneself and $500 for the other person as more satisfactory than the ratings given by another group to the outcome of $600 for oneself and $800 for the other person, revealing a preference for equal outcomes with less money over unequal outcomes with more.

Recent research also shows that decision makers may forego profits to avoid disadvantageously unequal allocations even in joint evaluation designs that allow for a direct comparison between alternative payoff schemes and, therefore, generally lead to profit maximization (Bazerman, Loewenstein, and White 1992). In these within subjects experiments, participants who must trade off higher absolute payoffs against the avoidance of disadvantageous inequality willingly choose the latter over the former (Garcia and Tor 2007; Garcia, Tor, and Gonzalez 2006; Garcia, Tor, Bazerman, and Miller 2005). In fact, these studies also reveal that where comparative evaluation plays a central role, decision makers forego profit even to obtain advantageous inequality (Garcia et al. 2006). Together, these findings imply defendants may reject comparatively unfair offers, although these offers are made by a prosecutor rather than a peer, and despite the fact that rejection does not penalize the prosecutor in any significant way.

Providing some suggestive evidence on comparative evaluation the criminal justice context specifically, Casper (1978) analyzed data from interviews with convicted felons. He found the felons' assertions of whether their sentences were heavier than those given to most others convicted of the same crime strongly correlated with their sense of whether their treatment was fair. A reanalysis by Casper et al. (1988) supported the above findings with respect to the important role played by comparative evaluation, independently of the effects of outcome severity and procedural justice. The apparent importance of comparative evaluation for post-verdict satisfaction indicates, in turn, that this factor may also impact defendants' ex-ante WTAP.

Finally, anecdotal evidence from cases such as that of Hayes, discussed above, also suggests a potential role for comparative evaluation in plea bargain decision making. Hayes recalled his interaction with Bagby:

He say, “if you don’t take the five years, I am going to indict you on the habitual criminal.” . . . I told him, “Look, man, you know, I have been in the State Reformatory one time and had one number on my back, you know,” and, I said, “there’s guys who have had six and seven numbers on their back and they never was tried on the habitual criminal. . . .” (cited in Stuntz 2005)

Apparently, Hayes compared Bagby's threat to charges brought against other criminal defendants who had far more felony convictions but "never was tried on the habitual criminal." Knowing the Kentucky prosecution did not use this charge for felons with many more "numbers," Hayes naturally found a plea offer in the shadow of the habitual criminal charge extremely unfair, on comparative grounds.
In sum, there is evidence for the role of comparative evaluation in related domains, but no direct findings in the plea bargaining context beyond anecdotes such as Hayes's. Studies 5-6 therefore sought to provide controlled experimental evidence on the impact of comparative evaluation on WTAP, showing the significant and independent role of this variable and providing new insights for the plea bargaining debate, which currently neglects comparative evaluation concerns altogether.

3.2. Study 5: The Comparative Evaluation of Plea Offers

Study 5 sought to extend the findings on the comparative evaluation of allocations to plea bargaining. We hypothesized that defendants' WTAP will be affected by comparative evaluation. This effect was expected to take place although the comparative standing of an offer would be irrelevant to a rational defendant. Specifically, we predicted that participants will become more likely to accept the bargain as its comparative advantage increases, even when the objective attractiveness of the offer remains fixed. Participants were thus predicted to consider offers more attractive when advantageously unequal than when disadvantageously unequal, compared to typical offers in similar circumstances.

Seventy four undergraduates from Eastern Michigan University and the University of Michigan (35 females and 39 males) volunteered at the library for a questionnaire survey. In a between-subjects design, participants read a scenario in which they were guilty of an offense, but their offer was either better or worse than those offered in similar situations:

Suppose a fire started because of your negligence and another student died as a result. You now face criminal charges, and there is a 50% chance you will be convicted in a trial and face a mandatory 2-year jail sentence. To avoid trial, however, the prosecutor has offered you a plea bargain. Although the prosecution typically offers a (3-month / 9-month) long jail sentence in similar circumstances, you have been offered a (longer / shorter) 6-month one.

Participants were then asked, “What would you do?,” and chose “Go to trial” or “Accept the plea-bargain.”

The results were consistent with the prediction. Among those whose plea bargain offer was shorter than the typical one, 67% accepted the plea with its 6-months jail sentence bargain. Only 42% of the participants whose offer was longer than the typical one, however, accepted the same plea bargain offer. This pattern was significant ($\chi^2=4.49, p<.05$). As predicted, therefore, and in line with both the anecdotal evidence and findings in other domains, participants' WTAP was significantly affected by comparative evaluation – although the objective attractiveness of the offer versus trial remained fixed – in clear contrast to common models of rational defendant behavior.

3.3. Study 6: Combining Substantive Fairness and Comparative Evaluation

The last study integrated the two phenomena we examine in this Article within a unified framework. With respect to comparative evaluation, this design allowed us, first, to generalize the effects to both substantively fair and substantively unfair offers; and, second, to study the effects of 3 levels of comparison – advantageous, similar, and disadvantageous. In addition, Study 6 also examined the relationship between substantive fairness and comparative evaluation, testing our hypothesis that these two analytically
independent constructs exert independent effects on WTAP even when simultaneously present.

Two groups of law students at University of Haifa in Israel (72 females and 48 males) completed a voluntary questionnaire in criminal procedure classes. In a 3 (substantive fairness: guilty / uncertain / not guilty) x 3 (comparative evaluation: better / similar / worse) mixed design, participants read a variant of the fire accident scenario used in Study 5:

Imagine a fire which started in your apartment while you were taking a class at the university caused the death of a neighbor. The police think you left turned on the space heater that caused the fire, but there are some difficulties in proving that. (You remember that you did leave / You don't remember whether you left / You remember you did not leave) the space heater turned on when you left the apartment.

Participants then read:

You are now standing on trial on a charge of negligent homicide. Considering all the evidence and circumstances, your conviction odds are 50%. If you are convicted, the judge will sentence you to a mandatory one-year jail sentence.

The next paragraph stated:

The prosecutor approaches you at the eve of trial and offers a (non-negotiable) plea bargain: You will plead guilty to a lesser offense, you will be convicted based on your admission, and you will be sentenced only to 3 months in jail.

At this point, participants in the 3 between-subjects substantive fairness conditions read 3 questions in which they had to choose between trial and plea. The counterbalanced offers only differed in their comparative evaluation. They asked:

What will you choose if the sentence offered by the prosecutor (3 months in jail) is (shorter than / similar to / longer than) the jail sentence typically offered by the prosecution in similar cases (6 months in jail / 3 months in jail / a month and a half in jail)?

We conducted a GEE analysis because of the correlated binary responses in this mixed design. As predicted, the contrast for the between-subjects factor of substantive fairness – 2 (guilty), -1 (uncertain), -1 (not guilty) – was significant ($\chi^2=4.85, p<.05$). The linear contrast for the within-subjects factor of comparative evaluation – 1 (better), 0 (similar), -1 (worse) – was also highly significant ($\chi^2=25.9, p<.001$) in accordance with our predictions. Finally, the interaction between substantive fairness and comparative evaluation was not significant ($\chi^2=1.83, p=.77$), suggesting that, even when both are present, the two variables exert independent effects on WTAP.

This pattern of results is clearly illustrated in Figure 2. Guilty condition participants showed higher rates of plea acceptance, with 59%, 44%, and 27% accepting the offer in the better than typical, similar to typical, and worse than typical within-subjects treatments, respectively. Uncertain and not guilty participants on the other hand,

---

31 See supra note 26.
showed lower acceptance rates, with 40% and 46% in the better than typical condition, 23% and 23% in the similar to typical condition, and 18% and 18% in the worse than typical treatments, respectively.

![Figure 2](image-url)

**Figure 2.** Deadly fire: percent accepting plea offer by levels of comparative evaluation

Altogether, the findings of Study 6 confirm our hypotheses and extend the basic findings of Study 5 regarding the effect of comparative evaluation, showing it to be independent of the impact of substantive fairness and consistent across different levels of this latter variable.

### 4. REVISITING THE PLEA BARGAINING DEBATE

The evidence examined here makes clear that defendants' evaluation of the fairness of plea offer terms plays an important role in their willingness to accept such offers. The extensive but largely theoretical plea bargaining literature, however, has paid little attention to the impact of either substantive fairness or comparative evaluation when evaluating plea bargaining policies. Most economic models of plea bargaining simply assume rationality and disregard fairness altogether (e.g., Grossman and Katz 1983; Reinganum 1988), thereby implying fairness considerations do not affect defendants'

---

32 Some exceptions can be found in the debate over *nolo contendre* and Alford Pleas. A *nolo contendre* plea ("a plea of no contest") allows defendant to accept a conviction without trial and without an explicit guilty plea. An *Alford* plea, (originated in the case of North Carolina v. Alford 400 US 25 [1970]) goes even further, allowing defendants explicitly to claim innocence while accepting the plea in order to minimize their risk. Bibas (2003), for one, argues that allowing such pleas encourages innocents who are reluctant to lie to accept a conviction (implying that, otherwise, innocents sometimes may reject otherwise attractive standard plea bargains).
behavior. Moreover, even those few scholars who considered potential differences between guilty and innocent defendants have relied on intuition instead of empirical evidence. Consequently, their conclusions often have been inaccurate.\footnote{For example, Bibas (2004: p. 2495) assumed innocent defendants are likely to be risk averse, accepting offers that guilty defendants would reject, in contrast with the extensive evidence discussed here. He also asserts that defendants who were intoxicated during their alleged crime and are thus uncertain about their guilt would be more easily persuaded by prosecutorial bluffing than their guilty counterparts to accept plea offers. However, our findings show this intuition is probably unfounded.} The following sections thus examine some of the tentative implications of our findings, showing they call both advocates and detractors of the rational defendant view to rethink and revise some of their long-standing positions.

4.1. The Cost of Innocence

Defendants have a preference for substantive fairness. This preference is not necessarily irrational; defendants may increase their utility by rationally rejecting substantively unfair plea offers, even at the cost of an increased risk of conviction at trial. However, the preference for substantive fairness bears important implications for legal policy. Most notably, innocent defendants who choose trial over plea will reject discounted offers that include sanctions lower than the expected sanction at trial. Inevitably, therefore, \textit{these innocents – as a group – will bear higher average penalties than guilty defendants} facing comparable conviction probabilities, who accept the discounted plea offers at significantly higher rates.

This somewhat counterintuitive "cost of innocence," where the preferences of innocents lead them collectively to fare worse than their guilty counterparts, is further increased by the "trial penalty." The trial penalty is the policy of imposing much harsher sentences at trial on defendants who contest the charges than on those who plead guilty (Wright 2005). This policy seeks to facilitate guilty pleas by guilty defendants, and would not cause disproportional harm to innocent defendants who hold no substantive fairness preferences (Church 1979; Easterbrook 1983). Ironically, however, the trial penalty disproportionately penalizes innocents, collectively, who reject on grounds of substantive fairness some offers their guilty counterparts accept.

Nevertheless, the negative, combined, effects of the cost of innocence and the trial penalty, may not apply to criminal prosecutions overall. Overall, prosecution cases against culpable defendants are likely to be stronger, on average, than those against innocents. Therefore, culpable defendants overall are probably still likely to fare worse than their innocent counterparts. Notwithstanding this caveat, however, there is a significant middle-range, where many guilty and innocent defendants face similar conviction probabilities. In this middle range, a taste for substantive fairness, whether rational or not, turns out to be costly indeed.

4.2. The Innocence Problem in a New Light

Much of the plea bargaining literature revolves around the "innocence problem" – that is, the possibility that innocent defendants accept discounted plea offers to avoid the risk of severe sentences following trial (e.g. Gazal-Ayal 2006; Wright 2005; Hessick and Saujani 2002; Schulhofer 1992; Scott and Stuntz 1992a). For many of the critics of plea bargaining this risk justifies a ban on plea bargaining (Alschuler 1968; Schulhofer 1992).
Yet, if innocent defendants tend to reject plea offers that guilty ones accept, the concern over the innocence problem may be exaggerated.

Ironically, while comforting scholars who fear that plea bargaining coerces wrongful guilty pleas, our findings should also cause supporters of plea bargaining to reevaluate their analyses. Many supporters argue the plea bargaining system only helps innocent defendants, offering them the option of "insurance" against the high stakes of imperfect trials (Scott and Stuntz 1992a; Easterbrook 1983). They argue that innocents would choose to plea bargain whenever the disutility from the expected sanction at trial is greater than the disutility from the bargained sentence and reject the plea otherwise (Lynch 1998, p. 2145; Easterbrook 1983). However, innocents' aversion to plea bargaining suggests this "insurance" mainly benefits guilty defendants and is of limited efficacy for innocents.

This conclusion is especially interesting, moreover, in light of current calls to replace the plea bargaining system with a system of simplified trials for all. Such calls mainly come from scholars who believe innocents should not be coerced to plead guilty (Alschuler 1983; Schulhofer 1984; Langbein 1979), while those who emphasize the need to reduce sentences imposed on innocent defendants usually support the plea bargaining system (Church 1979; Scott and Stuntz 1992a).

The picture becomes even more complicated once we consider the effect of egocentric biases on innocent and uncertain defendants. Innocents who overestimate their trial odds inevitably reject otherwise attractive plea bargains. This situation may not trouble opposers of plea bargaining by innocents (e.g., Alschuler 1981; Schulhofer 1992). However, the bias does pose a difficulty for supporters of innocents' freedom of choice (e.g., Church 1979; Easterbrook 1983; Scott and Stuntz 1992b), because it sometimes will lead innocent defendants mistakenly to reject pleas they would have accepted if they were not egocentrically biased.

Moreover, uncertain defendants will often refuse objectively attractive offers as well. For uncertain defendants who are in fact innocent, we are faced again with the cost of innocence and trial penalty problems, which should be especially troubling to supporters of plea bargaining freedom. Paradoxically, on the other hand, the bias of those uncertain defendants who are objectively guilty facilitates their choice of trial over plea, unwittingly leading them to what is, arguably, an appropriate venue for establishing their guilt. These uncertain, guilty but biased, defendants who refused a plea bargain and are later convicted, however, may not appreciate the truth-finding service provided by their criminal trial.

Nevertheless, plea aversion might serve to protect innocent and uncertain defendants if prosecutors can better ascertain guilt than trials. This often might be the case because prosecutors have better investigative tools, as well as access to information that is not transferred to juries. Since prosecutors rely heavily on guilty pleas, they are reluctant to prosecute defendants whom they expect to demand a jury trial (Gazal-Ayal 2006). Therefore, prosecutors may tend to divert their limited resources away from plea-averse defendants, among whom innocents are represented at a greater proportion than
among those defendants who are ready to plea, to reduce their risk of a jury trial.\textsuperscript{34} Prosecutors' incentives and innocents' plea aversion might thus combine to provide the latter with some unexpected protection.

4.3. Accounting for Acquittal Probabilities

One of the intriguing finding of our experimental tests was that the systematically different risk attitudes exhibited by innocent and guilty defendants, respectively also depend on acquittal probabilities. For medium to high probabilities, innocents are significantly more risk seeking than their guilty counterparts. This gap between the two defendant groups gradually narrows, then disappears, as the likelihood of acquittal diminishes.

This finding suggests that, despite their neglect of the impact of fairness considerations on defendants' behavior, the extant positions of both sides of the plea bargaining debate may still be relevant for a certain range of cases. When acquittal probabilities are low, substantive fairness concerns do not lead guilty and innocent defendants to significantly different plea acceptance behavior. For example, defendants who falsely confess during a police investigation might be correctly advised by their counsel their chances of acquittal at trial are slim. In such cases, they are likely to behave like an average defendant facing similar evidence, accepting the offer the prosecutor tailored for similarly situated guilty defendants. In fact, the exoneration data we analyzed in Part I, where 17 of the 19 guilty pleas of innocent defendants followed a false confession to the police,\textsuperscript{35} provide some support for this hypothesis.

4.4. Inequality in Plea Bargaining

The impact of fairness considerations on defendants' behavior may also have other fortunate consequences. For instance, the effect of comparative evaluation on defendants' willingness to accept plea offers may be used to curtail the inequality and arbitrariness of the sentences generated by plea bargaining. Scholars argue plea bargains lead to excessive diversity in sentencing, because prosecutors have much room to manipulate offers without risking rejections, given defendants' awareness of the trial penalty (Alschuler 1981; Stuntz 2004). This prosecutorial leeway, in turn, results in unequal treatment of similar cases. Yet, if defendants are averse to comparatively disadvantageous offers, an increase in the transparency of plea bargaining may curtail the diversity problem. For instance, if plea bargains were published, or if defense lawyers were to inform their clients about the common practices of prosecutors in similar cases, defendants would tend to reject comparatively harsh offers. As a result, prosecutors, who are interested in maintaining high guilty plea rates, would reduce disparity, arbitrariness and inequality in plea bargaining.

\textsuperscript{34} Prosecutors obviously have other reasons to refrain from charging innocent defendants, yet it would be wrong to assume they only choose those cases where innocence is least likely. For instance, prosecutors might prefer to dismiss a case against someone they know to be guilty, and proceed instead against someone whose innocence is possible, because the latter case is easier to prove or harder politically to dismiss given the nature of the alleged offense. In such cases, innocents' greater plea aversion may help tip the scales against proceeding with their prosecution.

\textsuperscript{35} Supra note 9 and the accompanying text.
While potentially beneficial, an increase in the transparency of plea bargain outcome may nevertheless generate its own difficulties. For one, transparency may exacerbate the effects of egocentric biases, which defendants exhibit with respect to substantive fairness judgments and are likely to impact comparative evaluations as well. In fact, some of the evidence from bargaining in civil disputes specifically records egocentric judgments in comparative evaluations (e.g. Babcock et al. 1995). Defendants with egocentrically biased judgments who obtain comparison information thanks to transparency, however, will tend to believe their offers are harsher than they truly are and reject offers that unbiased defendants would accept. A legal system wishing to avoid a choice between frequent biased rejections and a "race to the bottom" among prosecutors with increasingly discounted offers may therefore prefer to maintain the opacity of plea bargain outcomes.

4.5. The Behavioral and Experimental Study of Plea Bargaining

Our findings regarding the impact of substantive fairness and comparative evaluation on defendants' WTAP also highlight the importance of behavioral decision making for the legal analysis of plea bargaining. There is already evidence on the effect of another fairness-related variable – namely, procedural fairness – on defendants' decision making (e.g., Casper 1978; Casper et al. 1988; Houlden 1980). Furthermore, various hypothetical applications of behavioral phenomena to plea bargaining recently have appeared in the legal literature. Legal scholars have speculated, for example, about the influence of the self-serving bias and overconfidence on defendants (Bibas 2004), but failed to identify the likely interplay between these factors and fairness variables. These theoretical analyses also failed to identify the important role of substantive fairness more generally in shaping defendants' risk preferences, evidenced in our experiments, despite applying the construct of loss aversion, writ large, to defendants' decision making (Bibas 2004; Birke 1999). The largely speculative nature of the plea bargaining debate thus highlights the need for a systematic program for the empirical investigation of defendants' decisions making process in plea bargaining.

Finally, the study of plea bargaining would also be incomplete if it were to focus on defendants' decision processes alone. Other legal actors – most notably, judges, prosecutors and defense attorneys – play a central role in plea bargaining. For example, anecdotal evidence and ex-post surveys suggest that a defense attorney's recommendation plays a role in defendants' plea acceptance (Alschuler 1975; Bordens and Basset 1985; Casper et al. 1988). Prosecutors' decision behavior is also of primary importance (Alschuler 1968; Boylan and Long 2005), since plea bargaining is ultimately a strategic game: Prosecutors make offers they want defendants to accept. In some cases, judges can also reject the bargained sentence recommendation, a factor that both prosecutors and defendants must take into account when negotiating over pleas (Gazal-Ayal 2006; Scott and Stuntz 1992a). We believe the effects discussed here are substantial even after their diluted by other factors, as evidenced by the small number of guilty pleas in the database of wrongfully convicted defendants. Nevertheless, absent a comprehensive empirical study of the decision processes of defendants, prosecutors, defense attorneys, and possibly even judges, any formulation of plea bargaining policy and institutions will necessarily be incomplete.
References


