Article

The "Partial Innocence" Effect: False Guilty Pleas to Partially Unethical Behaviors

Personality and Social Psychology Bulletin
2025, Vol. 51(3) 335–356

DOI: 10.1177/01461672231185639
journals.sagepub.com/home/pspb
Sage

Stephanie A. Cardenas[1,2], Patricia Y. Sanchez[2], and Saul M. Kassin[2]

Abstract

Although research has focused on the "innocence problem," "partial innocence" may also plague individuals who plead guilty to crimes they did not commit, but that are either comparable, more severe, or less severe than their actual crimes. Using a high-stake experimental paradigm and an immersive role-playing paradigm, we examined the psychology of partial innocence. Students were randomly induced (or imagined themselves) to be innocent, guilty, or partially innocent of committing an academic transgression and then given the choice to accept or reject a deal to avoid disciplinary sanction. Across three studies (*Ns* = 88, 75, 746), partially innocent students pled to cheating nearly as often as guilty students and vastly more often than innocent students. Partially innocent students—not unlike guilty students—experienced greater feelings of guilt than did innocent students. In turn, these feelings of guilt, but not shame, were associated with taking responsibility for a range of transgressions not committed.

Keywords

guilt, shame, anger, emotion, deservingness, judgment, decision-making, plea bargaining

Received August 9, 2022; revision accepted June 12, 2023

According to combined data from the Innocence Project (n.d.) and the National Registry of Exonerations (n.d.), nearly 25% of over 3300 exonerees to date pleaded guilty to crimes they did not commit. This "guilty plea problem" includes defendants who are *partially* innocent—that is, innocent of the specific crime to which they pleaded guilty but culpable of a lesser, related, or prior offense (Zottoli et al., 2016).[1] For example, a defendant might be guilty of some counts, but not others (e.g., possessing, but not selling marijuana). When partially innocent (PI) defendants waive their Sixth Amendment right to trial, they do so not for reduced charges and sentences but rather for punishment equal to what they might have received if convicted of their actual crimes (Zottoli et al., 2016).

Cases of partial innocence pervade the American criminal legal system where over 95% of cases are resolved through guilty pleas (U.S. Sentencing Commission, 2022). For example, a field study of New York City residents convicted of a felony found that whereas 20% of defendants self-identified as completely innocent, twice as many (40.5%) self-identified as PI (i.e., having committed a similar but *different* offense to the one charged) (Zottoli et al., 2016). Some legal scholars argue that some PI defendants who are wrongfully convicted are likely known recidivists (Bower, 2008). This is because institutional biases make it more likely that police will stop and arrest the "usual suspects" and that prosecutors will charge them for crimes they did not commit that are consistent with their prior offenses. In other cases, an internal desire to take responsibility for analogous transgressions may motivate the accused's guilty plea. For example, Shanta Sweatt, an African American mother of two, pleaded guilty to possession of her ex-boyfriend's marijuana to "take ownership of what happens in the house" and to "protect her sons" (Yoffe, 2017). As a result of this admission, she lost her public housing and had to pay costly court fines and fees. Perhaps, given more time and fewer pressures, Shanta would have kept silent until she could speak with her lawyer and explain her situation.

Considering the immense direct and collateral consequences resulting from guilty pleas, it is important to assess

[1]Williams College, Williamstown, MA, USA
[2]John Jay College of Criminal Justice, City University of New York, New York City, USA

Corresponding Author:
Stephanie A. Cardenas, Department of Psychology, Bucknell University, One Dent Drive, Lewisburg, PA 17837, USA.
Email: scardenas@bucknell.edu

the role of the psychological state of partial innocence in plea decisions. Such research can broaden our understanding of the psychological phenomenon of partial innocence and it's effect on plea decision-making, as well as inform policies designed to mitigate false guilty pleas.

Innocence and Plea Bargain Decision-Making

Theorizing on the phenomenology of innocence that can put innocents at risk, Kassin (2005) suggested that factors at the criminal investigation stage, such as the naïve illusion of transparency and belief in a just world can lead innocent suspects to waive their Miranda rights (e.g., Kassin & Norwick, 2004; Scherr & Franks, 2015), to underreact, physiologically, to the accusation of guilt (Guyll et al., 2013), and even to confess under pressure expecting later to be exonerated (Perillo & Kassin, 2011). Confessions increase the likelihood that even innocent defendants would accept a plea offer (Perillo et al., 2014; Redlich et al., 2017). At the plea stage, innocents believe they will be viewed as such by attorneys and judges (Kassin, 2005) and endorse more optimistic beliefs about their probability of acquittal (Gregory et al., 1978; Tor et al., 2010)—even when both groups face objectively similar odds. This literature highlights the importance of examining to what extent a phenomenology of partial innocence may also uniquely characterize the behaviors, emotions, and cognitions of individuals guilty of wrongdoing different from the misconduct of which they have been accused.

A Phenomenology of Partial Innocence

Research shows that people interacting with the criminal legal system desire a fair adjudication process, respectful treatment, and the opportunity to have their voices heard (Tyler, 2000). When considering plea offers, innocent and guilty defendants alike reject unfair offers framed as comparatively worse than those proffered to other similar defendants (Tor et al., 2010). This attention to fairness may be rooted in their belief in a just world (Lerner, 1980) in which good actions deserve rewards and bad actions deserve punishment (Feather, 1999).

We propose that PI defendants may believe that they deserve punishment for being "culpable of something" and that feelings of guilt may partly account for their disproportionately high plea rates. In support of this hypothesis, basic attribution research shows that people can attribute causality between their negative actions and punishment they receive even when *no* causal connection exists (Callan et al., 2014). Research on immanent justice reasoning suggests that the belief in a just world can lead people to draw incorrect causal connections between their unethical behaviors and subsequent random bad breaks (Lerner, 1980).

Once punishment is perceived as deserved, it can trigger feelings of guilt and shame (see Feather et al., 2011)—non-mutually exclusive moral emotions that can motivate behaviors (or "action tendencies") intended to alleviate these feelings (Callan et al., 2014; Feather, 2006; Feather et al., 2011; Tangney et al., 2011) and which may represent different paths to guilty pleas. For example, whereas guilt-proneness and state feelings of guilt are associated with adaptive behaviors such as perspective-taking and empathizing, shame-proneness and feelings of shame instead tend to be associated with maladaptive behaviors such as externalizing blame, interpersonal distancing, defensiveness and expressions of self and other-oriented anger (e.g., "humiliated fury"; Tangney et al., 2007). These discrepant behaviors are the result of differences in the target of focus among people feeling guilty versus ashamed.

People feeling guilty tend to focus on the ways in which their "bad behavior" negatively affects others. This empathic response then motivates them to "right the wrong[s]" of their transgression leading them to engage in reparative actions such as confessing, apologizing, and compensating for the consequences of their wrongdoing (Cialdini et al., 1973; Tangney et al., 2007, p. 6). Indeed, people feeling guilty also become more compliant, for example, by fulfilling the injured party's compensatory requests to donate time or money (for a meta-analysis, see Boster et al., 2016). They may even comply with non-compensatory requests from individuals other than the injured party that mitigate guilty feelings while doing nothing to repair prior harms (Carlsmith & Gross, 1969; Freedman et al., 1967). In support of a similar role of guilt in plea contexts, real and mock defendants retrospectively attribute their guilty pleas to feelings of remorse and a desire to take responsibility (Bordens & Bassett, 1985; Redlich & Shteynberg, 2016; Wilford et al., 2018).

People feeling ashamed instead tend to have an egocentric focus on the "bad self" (as opposed to their "bad behaviors") and are thus less likely to reflect on the harmful consequences of their wrongdoings for others. This decreased empathic concern is joined by attempts to deny their wrongdoing and to escape from the shame-inducing situation. Research consistently links shame with a variety of risky, illegal, and problematic behaviors (see Tangney et al., 2007). Thus, while feeling *guilty* may predict increased responsibility-taking in the form of admissions of guilt, feeling *ashamed* may predict the opposite tendency (for a similar argument, see Boster et al., 2016, p. 62).

Overview of the Current Study

Our goals were to understand the psychological state and effects on plea decision-making of individuals who are PI of the transgressions for which they are accused. In Study 1 ($N = 88$), we employed a three-group between-subjects design using adapted versions of the cheating paradigm (Russano et al., 2005) and Dervan and Edkins's (2013) and Perillo et al.'s (2014) high stakes deception-based plea simulation

paradigms. Participants were experimentally induced to cheat (or not induced to cheat) on an academic task. To induce partial innocence, we randomly assigned half of participants who cheated to receive an accusation of having cheated on a *separate*, but comparable, task of which they were factually innocent. Guilty participants were instead accused of cheating on the *same* task on which they had cheated. Following the accusation, a "project supervisor" offered all participants a plea deal in which they could avoid academic sanction in exchange for admitting guilt, agreeing to 40 hours of community service, and waiving their right to a hearing before an academic disciplinary committee. After rendering a plea decision, participants reported on their feelings of guilt and shame, and the decision-making experience, more broadly. In Studies 2 ($N = 75$) and 3 ($N = 746$), participants imagined themselves in an immersive role-playing version of the high-stakes paradigm of Study 1. To understand the phenomenology of partial innocence more fully, Study 3 distinguished between two novel manipulations of partial innocence—wherein participants were accused of a transgression that was either *more* or *less* severe than their actual misconduct. Though the current work is primarily exploratory, we set out to test three main predictions.

Hypothesis 1. We expected that Partially innocent (PI) participants would plead guilty at higher rates than innocent participants.

Hypothesis 2. We anticipated that PI participants would self-report greater feelings of guilt than Innocent participants.

Hypothesis 3. Finally, we predicted that feelings of guilt would be positively correlated with acceptance of the plea offer.

Study 1

Method

Participants. A total of 168 undergraduates from an urban commuter college in the northeast were recruited for a paid (US$30) in-person laboratory experiment on "Social Perceptions." Students had to be registered for classes to enable the threat of academic sanction. To reduce the risk of enrolling students who might find the high-stakes paradigm too stressful (Perillo et al., 2014), participants were required to be 18 years or older, have a grade point average (GPA) above 2.0, and pass an online anxiety screener ($N = 34$ excluded; Beck et al., 1988). Of the remaining 131 participants, we were able to schedule 107 for the laboratory phases: 19 were excluded from final analyses (2 innocent, 7 guilty, 10 PI) for failing to cheat despite inducement ($n = 9$), cheating without inducement ($n = 1$), reporting suspicion prior to debriefing ($n = 4$), terminating the session due to significant distress ($n = 4$), and/or for failing to make a plea decision within the allotted time of 6 min ($n = 6$, evenly split across conditions).

After exclusions, the final sample consisted of 88 participants: 28 Guilty, 31 Innocent, and 29 PI (ages: 18 to 32, $M = 20.86$, $SD = 3.68$; 73.9% females). Overall, 39.8% identified as Hispanic/Latino, 28.4% as Black/African American, 23.9% as Asian, 17% as White, and 3.4% as Other. We based our sample size primarily on resource constraints involved with designing a cost and time-intensive paradigm (Läkens, 2022). We performed a sensitivity power analysis in G*Power 3.1 for the feelings of guilt variable (H3), comparing the mean of the Accept versus Reject Plea Deal groups with an independent samples *t*-test, assuming a one-tailed test and an alpha of .05. A sample of this size ($n = 83$, as 5 participants did not provide emotion ratings) would provide 80% power to detect an effect of Cohen's $d = .68$ and 99% power to detect an effect of Cohen's $d = .1.08$. For reference, among the eight studies that measured state feelings of guilt, Boster et al.'s (2016) meta-analysis found an effect size of $d = 1.06$ for the association between feelings of guilt and increased compliance, broadly defined.

Materials and Procedure. All study materials and analysis code are available on the Open Science Framework (OSF): osf.io/v2r9e. Data for Study 1 (osf.io/7vjn4), Study 2 (osf.io/fdme4), and Study 3 (osf.io/tbz2p), as well as the corresponding codebook are also available (osf.io/8uaz7). We report all manipulations, measures, and exclusions for all studies. For Study 3, all hypotheses, sample sizes, materials, procedures, inclusion and exclusion criteria, data preparation and statistical analysis code were pre-registered and are available at osf.io/7bn8. We report all pre-registered analyses in the main body of the manuscript. There were no deviations from the pre-registered confirmatory analysis plan.

Pre-Experiment Questionnaires. To examine the influence of partial innocence on plea decisions above and beyond the effect of individual differences across related constructs and to increase statistical power and precision in our models, we first administered measures to control for trait interrogative compliance, trait deservingness of bad outcomes, guilt and shame proneness, just world beliefs (general and personal), and self-esteem (αs = .62-.83). However, because there were no bivariate correlations between these measures and plea decision, we do not discuss these further. See OSF (osf.io/uhq25) for a full description of each measure and means, standard deviations, and correlation matrices split by condition (Guilty: osf.io/y6js5; Innocent: osf.io/u9dzk; PI: osf.io/t9v2b).

Cheating Paradigm Task. Next, participants were scheduled for a laboratory session. Three trained research personnel helped run each session: a female experimenter to guide participants through the tasks; a female confederate to deliver the manipulation; and a male "project supervisor" to deliver the plea offer.[2] Each laboratory session took about 1 hr and comprised five highly-scripted phases: a rapport-building

phase between participants and their confederate partner, a culpability induction, an accusation, a plea offer, and a post-decision questionnaire.

Laboratory phases were covertly video recorded. Sessions began with the experimenter explaining to participants-confederate pairs that the researchers were interested in understanding how individuals and teams differ in their problem-solving capabilities. As such, participants were given a few minutes to become acquainted with each other. This rapport-building phase was included to incentivize participants to comply with the confederate's requests for help in a subsequent task where collaboration was forbidden (described below). To further incentivize cheating, the experimenter noted the possibility of an additional performance-based compensation (up to $15 USD) on two occasions. Next, participants solved two problem sets individually and two collaboratively with the confederate. In one third of conditions, the confederate complied with instructions to work alone on the both individual tasks (*Innocent* condition); in two-thirds of conditions, the confederate induced participants to cheat by sharing and soliciting information on the second individual task (*Guilty* and *PI* conditions). Participants then rated their impressions of their partner and the experimenter while the experimenter ostensibly scored their responses.

Accusation and Manipulation. After pretending to score their responses, the experimenter returned, stated there was a problem and escorted the confederate to another room. Upon her return a few moments later, she accused the participant of cheating on either "Individual Task 1" (IT1) or "Individual Task 2" (IT2). Participants who were both induced and accused of cheating on IT2 comprised the *Guilty* group and those accused of cheating on IT1 but induced to cheat on IT2 comprised the *PI* group. To ensure that PI participants understood they were accused of cheating on IT1, not IT2, the experimenter demonstrated the problematic task on a blank copy of the worksheet. This precautionary measure was repeated twice (once by the project supervisor) and two other times verbally. Participants not induced to cheat on any task were in the *Innocent* group regardless of accusation type. After the accusation, the experimenter left to ostensibly contact her superiors.

Plea Deal and Decision. Five minutes later, the "project supervisor" arrived and informed participants that the professor in charge of the study considered cheating a form of academic dishonesty and planned to file charges with the college's "Academic Integrity Board." The supervisor then presented a deal: in exchange for pleading guilty and completing 40 hours of community service, which would appear in the student's internal record only, the professor would not file charges. In contrast, refusal to accept the plea would lead to a hearing at which the participant would testify, the outcome of which was unknown (to allay fears of the most severe possible penalty, participants were told that they would not be expelled, regardless). Participants rendered a plea decision by signing the waiver form. The project supervisor exited the room ostensibly to speak with the experimenter.

Self-Reports and Debriefing. Next, the experimenter entered the room and probed participants for suspicion by asking the participant a series of questions: "*How are you feeling? What are you thinking? Why is that? Can you tell me a little more about that? What did you decide? Why did you accept/reject the offer?*" The experimenter then assured participants that they were not in trouble, and promised to explain the study after participants completed a final questionnaire. Five participants had to be fully debriefed at this time (two Innocent and three PI). Those who remained ($n = 83$) completed a post-decision questionnaire where they reported feelings of "guilt" and "shame" using 1 item each randomly presented as part of the 20-item Positive and Negative Affect Scale (PANAS; Watson et al., 1988) on 5-point Likert-type scale (1 = *very slightly or not at all* to 5 = *extremely*), as well as other exploratory items reported in the Supplemental Materials. A subset of participants ($n = 75$) completed manipulation checks (explained below). All were then fully debriefed. A complete description of the experimental protocol and script followed by all research personnel is available at osf.io/hjnw8.

Analytic Strategy. We present results from one-way ANOVAs with a Huber-White correction for unequal variances, Welch's *t*-tests for unequal variances (Delacre et al., 2017), planned Games-Howell post hoc comparisons, bias-corrected Cohen's *d* effect sizes according to Hedges and Olkin's (1985), bias-adjusted pseudo R^2 Nagelkerke effect sizes, and 95% confidence intervals.

Results

Preliminary Analyses. Participants did not significantly differ in age, GPA, anxiety, or any personality traits and beliefs across conditions ($ps > .36$). Plea responses did not differ according to confederate, experimenter, or project supervisor ($ps > .07$). Prior to a full debriefing, participants identified the individual task (IT1 or IT2) on which they had been accused of cheating. Most participants (94.7%)—including all PI participants—reported the task accurately. When asked to identify the task on which they had *actually* cheated, 58.3% of Guilty participants, compared with only 20.7% of PI participants, accurately reported cheating on IT2. Across both groups, incorrect reporting consisted primarily of denials that they cheated on either task. Because we previously confirmed that all Guilty and PI participants had cheated, we interpret this inaccurate reporting as a reluctance on behalf of participants to incriminate themselves on paper, especially prior to a full debriefing. Furthermore, PI participants reported ratings of self-perceived culpability (1 = *completely*

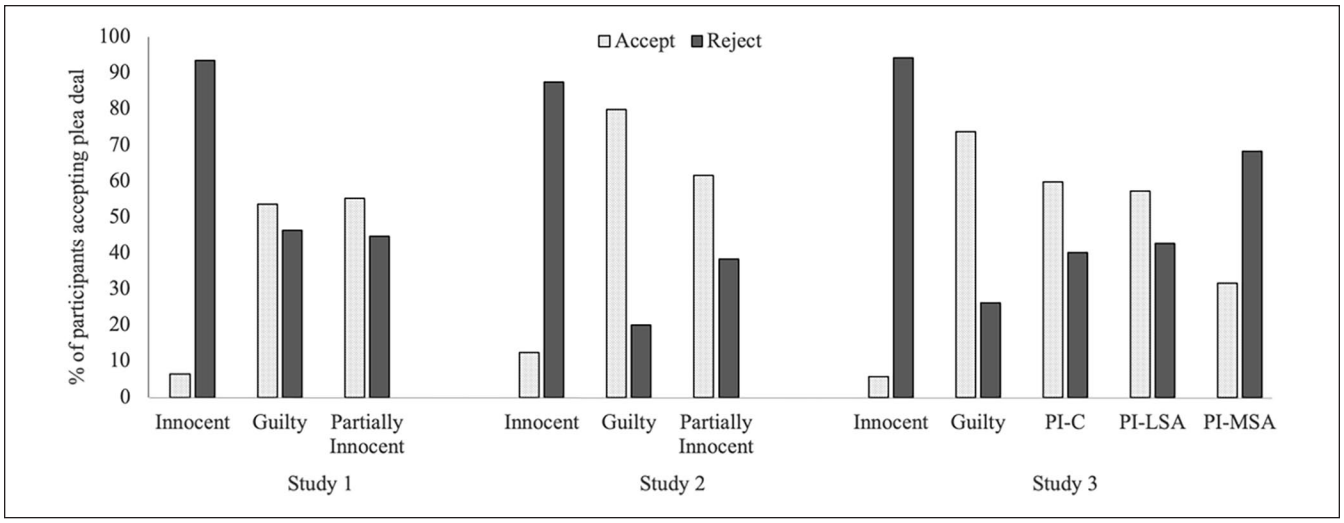

Figure 1. Percentage of Participants Who Rejected and Accepted the Plea Offer in Each Condition.
Note. PI-C = Partially Innocent—Comparable Accusation; PI-LSA = Partially Innocent—Less Severe Accusation; PI-MSA = Partially Innocent—More Severe Accusation.

Table 1. Logistic Regressions Predicting Likelihood of Plea Decision by Culpability.

Effect	*B*	*SE*	*Wald* z	*p*	Odds ratio	95% CI for odds ratio	
						Lower	Upper
Study 1							
Constant	0.14	0.38	0.38	.705	1.15	0.55	2.46
PI-C vs. Innocent	−2.82	0.82	−3.42	<.001	0.06	0.01	0.25
PI-C vs. Guilty	0.06	0.53	0.12	.903	1.07	0.37	3.05
Study 2							
Constant	0.47	0.40	1.66	.243	1.60	0.73	3.65
PI-C vs. Innocent	−2.42	0.74	−3.28	.001	0.09	0.02	0.34
PI-C vs. Guilty	0.92	0.64	1.43	.153	2.50	0.73	9.44
Study 3							
Constant	0.39	0.16	2.40	.016	1.48	1.08	2.06
PI-C vs. Innocent	−3.19	0.38	−9.86	< .001	0.04	0.02	0.08
PI-C vs. Guilty	0.64	0.25	−.260	.009	1.89	1.17	3.07
PI-C vs. PI-LSA	−0.10	0.24	−2.96	.675	0.91	0.57	1.44
PI-C vs. PI-MSA	−1.17	0.25	−6.96	< .001	0.31	0.19	0.50

Note. CI = confidence interval; PI-C = Partially Innocent—Comparable Accusation; PI-LSA = Partially Innocent—Less Severe Accusation; PI-MSA = Partially Innocent—More Severe Accusation.

innocent to 7 = *completely guilty*) that were not significantly different from those of Guilty participants, $t(51.9) = -1.74$, $p = .199$, $d_{cohen} = -0.48$ [−1.14, 0.06]), but that were significantly greater than ratings in the Innocent condition, $t(40.4) = 2.51$, $p = .041$; $d_{cohen} = 0.65$ [0.12, 1.21]. Guilty participants reported feeling more culpable than those who were Innocent, $t(80) = 4.42$, $p < .001$, $d_{cohen} = 1.22$ [0.66, 1.9]). Collectively, these results demonstrate that PI participants understood they had been accused of cheating on a task *different* from the one they had actually cheated on and were more reluctant than Guilty participants to incriminate themselves again.

Confirmatory Analyses

H_1: Culpability Predicts Plea Decision. Turning to our most critical dependent measure, Figure 1 shows the plea rates broken down by condition. Results support H_1: in contrast to participants who were innocent, very few of whom agreed to plead guilty (6.5%), PI participants accepted the plea at a significantly higher rate—not different from those who were guilty (55.1% and 53.5%, respectively). See Table 1 for logistic regression statistics.

H_2: Culpability Is Associated With Feelings of Guilt and Shame. On self-reported feelings of guilt, a one-way ANOVA

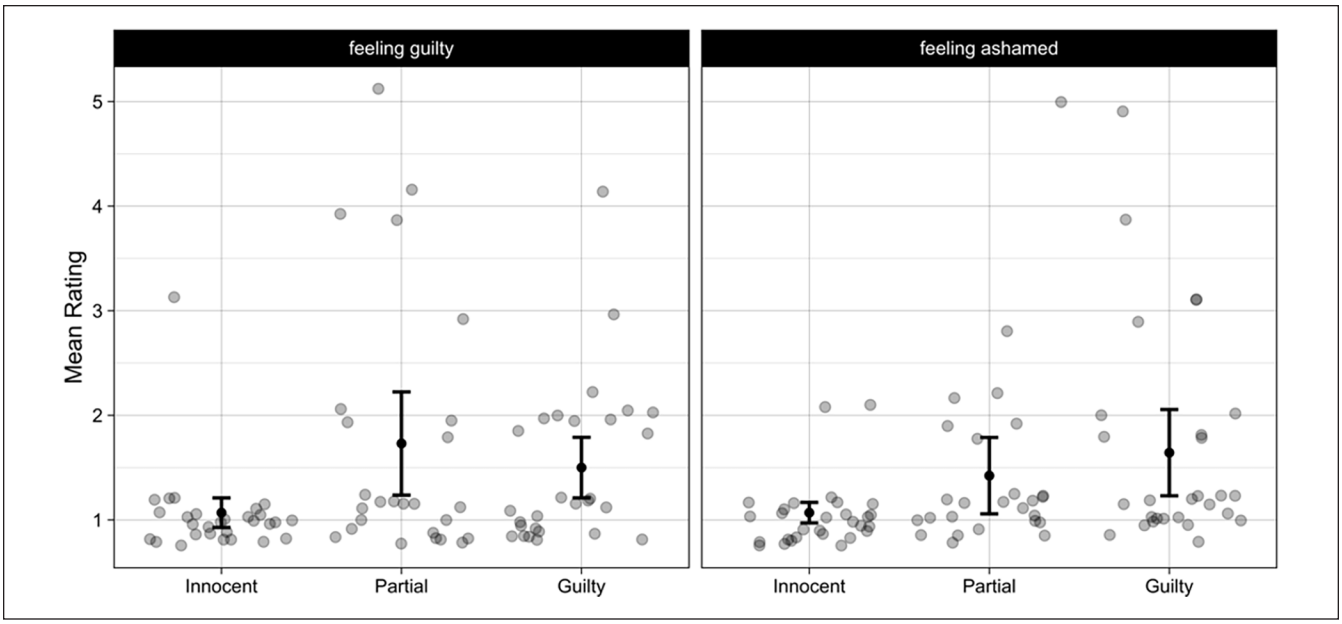

Figure 2. Distribution of Mean Ratings of Guilt and Shame in Study 1.
Note. Dot plot: Error bars represent 95% CIs. Feelings of guilt and of shame measured using a single item each rated as part of the 20-item PANAS (1 = *very slightly or not at all* to 5 = *extremely*). CI = confidence interval; PANAS = Positive and Negative Affect Scale.

revealed a significant effect of condition, $F(2, 80) = 6.28$, $p = .003$, $\eta^2 = .10$ [.00, .23]: feelings of guilt were significantly *greater* among PI participants compared with Innocent participants, $M_{diff} = 0.66$; $t(29.2) = 2.66$, $p = .033$, $d = 0.72$ [0.27, 1.12], and among Guilty compared with Innocent participants, $M_{diff} = 0.43$; $t(39.3) = 2.75$, $p = .024$, $d = 0.72$ [0.23, 1.29], but did not significantly differ between PI and Guilty participants, $M_{diff} = 0.23$ [-.44, .91]; $t(40.8) = .83$, $p = .68$, $d = 0.23$ [−0.38, 0.71]).

Next, we examined whether self-reported feelings of shame differed across conditions. The ANOVA indicated a significant effect $F(2, 80) = 5.23$, $p = .007$, $\eta^2 = .08$ [.00, .21]: feelings of shame were not significantly different among PI participants compared with either Innocent, $M_{diff} = 0.35$; $t(28.7) = 1.93$, $p = .15$, $d = 0.53$ [−0.07, 0.93], or Guilty participants, $M_{diff} = -0.22$; $t(51.6) = -0.82$, $p = .69$, $d = -0.22$ [−0.75, 0.34]. Only those in the Guilty condition felt significantly more ashamed than those who were Innocent, $M_{diff} = 0.57$; $t(30.1) = 2.78$, $p = .024$, $d = 0.73$ [0.39, 1.11].

Figure 2 shows the distribution of reported feelings. Visual inspection of the distributions across conditions reveals greater dispersion in the Guilty and PI conditions. Notably, among participants reporting emotions, 34.61% ($n = 9$ of 26) and 39.28% ($n = 11$ of 28) of PI and Guilty participants, respectively, reported feelings of guilt greater than 1 (*not at all feeling this way*) compared with just 3.44% ($n = 1$ of 29) in the Innocent condition; similarly, 35.71% ($n = 7$) and 26.92% ($n = 10$) of PI and Guilty participants, respectively, reported feelings of shame >1 compared with just 6.8% ($n = 2$) of Innocent participants.[3]

H_3: Planned Mediation Paths to Guilty Pleas. To examine the two routes by which moral emotions—feelings of guilt or shame—could predict plea decisions, we first confirmed the unconditional effect of each emotion on plea decision.[4] We then fit a parallel mediation model with the sem() function in the *lavaan* package for *R* (Rosseel, 2012).[5] This analytic plan allowed us to control for the *Innocent* versus *Guilty* dummy-coded contrast variable (X2) and feelings of shame—a related potential simultaneous mediator. To test the null hypothesis that the relative indirect effects through guilt and shame mediators do not significantly differ from each other, we specified a contrast for the two relative indirect effects, subtracted these from each other, and assessed whether the difference reached statistical significance. Figure 3 shows the results of this model along all paths.

As expected, increased feelings of *guilt* (b_1 path: $b = 0.15$ [0.07, 0.25], $p < .001$), but also of *shame* (b_2 path: $b = 0.13$ [0.005, 0.23], $p = .016$), were associated with increased plea acceptance, even controlling for culpability condition. As expected, the indirect effect of the *Innocent* versus *PI* contrast through feelings of *guilt* was significant (a_2*b_1: $b = 0.10$ [0.02, 0.21], $p = .037$). Conversely, the indirect effect of the *Innocent* versus *PI* contrast through feelings of *shame* was not significant (a_4*b_2: $b = 0.05$ [−0.001, 0.12], $p = .134$). To formally test which route better predicted plea decision, we compared the two indirect effects through guilt and shame and found that these did not significantly differ ($b = 0.06$ [−0.04, 0.18], $p = .335$).

Unexpectedly, neither indirect effect of the *Innocent* versus *Guilty* contrast through feelings of *guilt* (a_1*b_1: $b = 0.066$ [0.01, 0.07], $p = .056$) or *shame* (a_3*b_2: $b = 0.075$ [0.002,

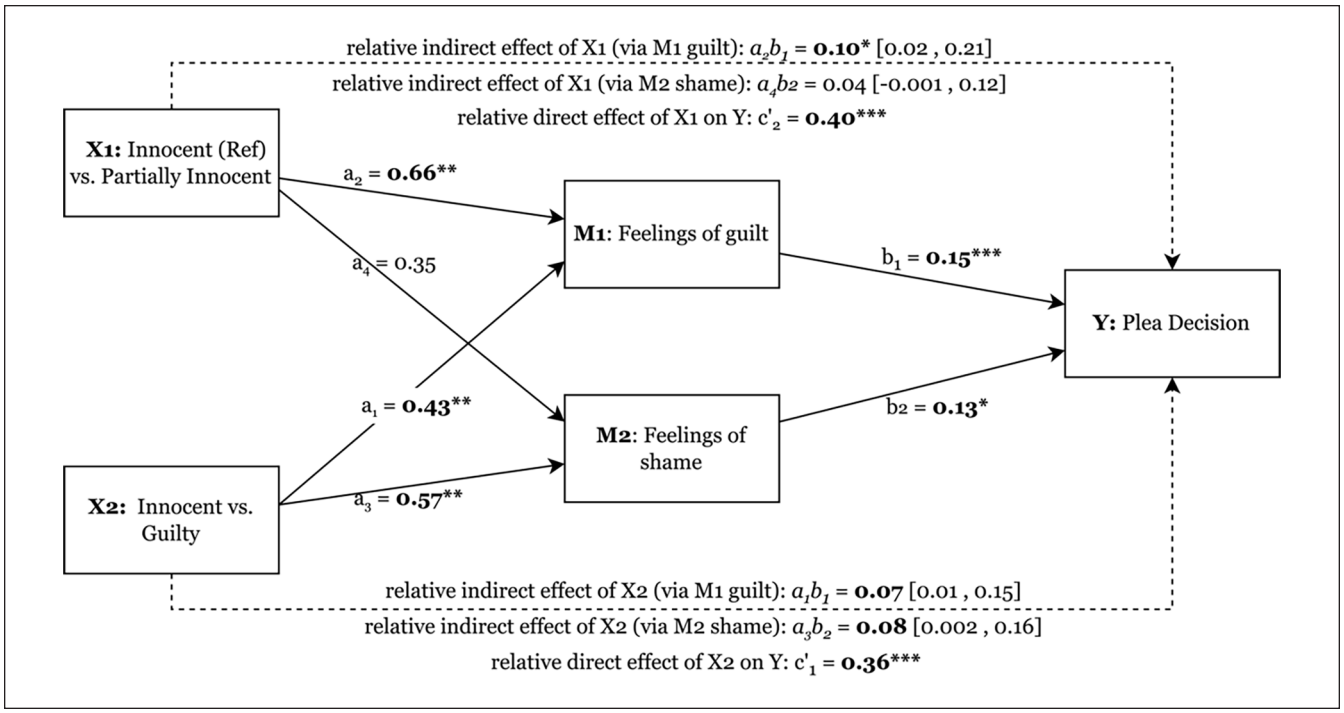

Figure 3. Parallel Mediation Model of Culpability Predicting Plea Decision Mediated by Feelings of Guilt and Shame in Study 1.
Note. Results of a parallel mediation analysis examining the relative indirect effects of condition contrasts (X1 and X2) on plea decision (Y) through state feelings of guilt (M1) and shame (M2). Unstandardized estimates are displayed with 95% confidence intervals. Bolded only, $p < .07$; $n = 83$. Reference Group: Innocent: = 1, All others = 0; Plea Decision: Accept = 1, Reject = 0. Estimated coefficients are based on bootstrapping procedure with 10,000 bootstrap samples. See Supplemental Figure 1 for the same model with Partially Innocent as the Reference Group.
*$p < .05$. **$p < .01$. ***$p < .001$.

0.16], $p = .067$) reached statistical significance. As expected based on results reported above, testing the indirect effect of the *Partially Innocent* versus *Guilty* contrast through feelings of *guilt* ($b = -0.04$ [−0.13, 0.05], $p = .433$) and *shame* ($b = 0.03$ [−0.04, 0.12], $p = .460$) in a model with PI as the reference group showed that neither effect reached statistical significance.

In sum, this analysis tentatively supports the idea that relative to being completely innocent, merely being culpable of *something* comparable—even when that *something* was not the target of the accusation—was associated with greater feelings of *guilt* and *shame*; and the more guilty, but perhaps not ashamed, individuals felt, the greater the frequency with which they accepted the plea offer. Of note, three reasons should give us pause when interpreting these results: (a) the indirect effects through feelings of guilt and shame did not significantly differ, (b) results are based on a small sample of participants ($n = 83$), and (c) emotion measures were collected *after* participants rendered a plea decision.

Discussion and Introduction to Study 2

PI participants—in stark contrast to innocent participants—pled guilty to an offense they did not commit and did so nearly as often as guilty participants. Results suggest that many PI participants experienced moderate feelings of guilt—not unlike those felt by guilty participants, and significantly greater than those reported by innocent participants ($d = 0.72$ [0.27, 1.12]). Feeling guilty, in turn, was associated with pleading guilty. In Study 2, we sought to conceptually replicate our findings with an online student sample that was asked to *imagine* themselves facing similar circumstances to those encountered by Study 1 participants. This conceptual replication served as a proof of concept for a larger, more immersive role-playing replication and extension study.

Hypothesis 4: We also sought to examine whether judgments of deservingness would be more similar between PI and Guilty participants compared with Innocent participants.

Study 2

Method

Participants. We recruited a convenience sample of 79 students ($M_{age} = 19.95$ years, $SD = 1.12$; 30 male, 42 female, 3 non-binary) from the research participant pool at a small (<2,500) U.S. liberal arts college for an online study on academic misconduct. After excluding participants who failed two or more attention checks, 75 remained. A sensitivity analysis indicated that an independent sample's t-test for pairwise comparisons with 25 participants per group (n= 50 per

comparison pair) would be sensitive to reliably detect effects larger than or equal d_{cohen}= 0.71 with 80% power (alpha = .05, one-tailed). The current sample was sufficient to detect effects at least as large as those observed in Study 1 for H_2 (d_{cohen}=0.76 [0.21, 1.30]) and H_3 (d_{cohen}=0.94 [0.51, 1.44]).[6]

Materials and Procedure. Participants read a description of the scenario encountered in Study 1 and imagined themselves in that situation. After providing their first name at the beginning of the survey, their name was embedded throughout the scenario to increase realism and perspective-taking. Participants were randomly assigned to learn that they were *Innocent* (n = 24), *PI* (n = 26), or *Guilty* (n = 25) of academic misconduct and were presented with a plea offer resembling the one shown to participants in Study 1 on a letterhead pertaining to their institution. Prior to rendering a plea decision, participants were asked to describe how they might feel about their behavior during the research session, being accused of academic misconduct, and having to render a decision. Next, they completed the PANAS, rendered a plea decision, explained their decision, and then evaluated various attributes of the case—that is, evidence strength, evidence influence on plea decision, and their estimated likelihood of conviction (see Supplemental Table 4 for exploratory analyses relating to plea judgments). Then, participants rated how deserving they were of their predicament and the fairness of the outcome (α = .89) and completed the General Just World Belief scale (α = .79). See OSF for a correlation matrix of Study 2 variables split by condition (Guilty: osf.io/pmvjg; PI: osf.io/swxvm; Innocent: osf.io/j3b6q). Figure 4 shows the distribution of emotion, deservingness, and evidence ratings.

Results

Confirmatory Analyses

H_1: Culpability Predicts Plea Decision. As in Study 1, a binary logistic regression supported H_1: PI participants (61.53%) were significantly more likely to accept the plea deal than Innocent (12.5%) participants, but not significantly more likely than Guilty participants (80%). See Table 1.

H_2: Culpability Predicts Feelings of Guilt and Shame. Replicating Study 1, a one-way ANOVA demonstrated a significant effect of condition on feelings of guilt, $F(2, 72) = 10.32$, $p < .001$ $\eta^2 = .16$ [.03, .31]: feelings of guilt were significantly *greater* among PI participants compared to Innocent participants, $M_{diff} = 1.08$; $t(35.9) = 3.58$, $p = .003$, $d = 0.99$ [0.49, 1.59], and among Guilty compared to Innocent participants, $M_{diff} = 1.02$; $t(36.2) = 3.61$, $p = .003$, $d = 1.01$ [0.53, 1.6], but did not significantly differ between PI and Guilty participants, $M_{diff} = -0.06$; $t(48.8) = 0.17$, $p = .99$, $d = 0.05$ [−0.51, 0.59].

Replicating Study 1, an ANOVA demonstrated a significant effect of condition on feelings of shame, $F(2, 72) = 7.51$, $p = .001$ $\eta^2 = .12$ [.01, .26]: Innocent participants felt significantly less ashamed than PI participants, $M_{diff} = 0.78$; $t(36.5) = -3.03$, $p = .012$, $d = -0.83$ [−1.36, −0.32], and Guilty conditions, $M_{diff} = -0.83$; $t(34.2) = -3.08$, $p = .011$, $d = -0.86$ [−1.42, −0.38]. Feelings of shame did not significantly differ between PI and Guilty participants, $M_{diff} = -0.06$; $t(34.2) = 0.17$, $p = .99$, $d = -0.03$ [−0.6, 0.53].

H_3: Feelings of Guilt and Shame Are Associated With Plea Decision. Consistent with Study 1, increased feelings of guilt ($b = 0.72$, $SE = 0.24$, $p = .002$) and shame ($b = 0.72$, $SE = 0.26$, $p = .006$) were associated with plea acceptance. Adding culpability condition to each model—for feelings of guilt, $\chi^2(2) = 17.94$, $p < .001$, and shame, $\chi^2(2) = 19.48$, $p < .001$,—significantly improved model fit.

H_4: Culpability Predicts Judgments of Deservingness. A one-way ANOVA demonstrated a significant effect of condition on deservingness, $F(2, 72) = 37.52$, $p < .001$ $\eta^2 = .40$ [.22, .53]: PI did not feel significantly more deserving than Guilty participants, $M_{diff} = 0.24$; $t(47.9) = 0.76$, $p = .730$, $d = 0.21$ [−0.35, 0.78]. But, as expected, on average, PI, $M_{diff} = 1.77$; $t(6.76) = 33.1$, $p < .001$, $d = 1.86$ [1.34, 2.63], and Guilty participants, $M_{diff} = 1.53$; $t(6.74) = 34.7$, $p < .001$, $d = 1.88$ [1.33, 2.73], each felt more deserving of punishment than Innocent participants.

Discussion and Introduction to Study 3

Study 2 replicated our initial laboratory experiment in a low-stakes, online role-playing paradigm. PI participants reported feelings of guilt not unlike those of Guilty participants, but greater than those in the Innocent condition ($d = 0.99$ [0.49, 1.59]). Feeling guilty and ashamed was associated with pleading guilty, even when controlling for actual culpability. Interestingly, results suggest that students asked to *imagine* themselves as culpable of academic misconduct reported more extreme moral emotions than students who actually engaged in academic misconduct, a point we return to in the General Discussion. Despite this discrepancy in levels of absolute reported emotions, a similar pattern of relative differences in moral emotions and plea decisions across culpability conditions emerged across Studies 1 and 2.

In Study 3, we sought to examine the boundary conditions of these effects. Previously, we operationalized "partial innocence" narrowly by inducing participants into an ethically comparable transgression. Study 3 advanced its predecessors in three ways. First, in addition to the Guilty, Innocent and PI conditions we added two variations of partial innocence—those accused of *less* severe transgressions and those accused of *more* severe transgressions. An individual accused of a *less* severe transgression than the one they committed should believe themselves deserving of punishment, feel guilty, and plead guilty at rates similar to those accused of comparable transgressions. Conversely, an individual accused of a *more* severe transgression than the one they committed should,

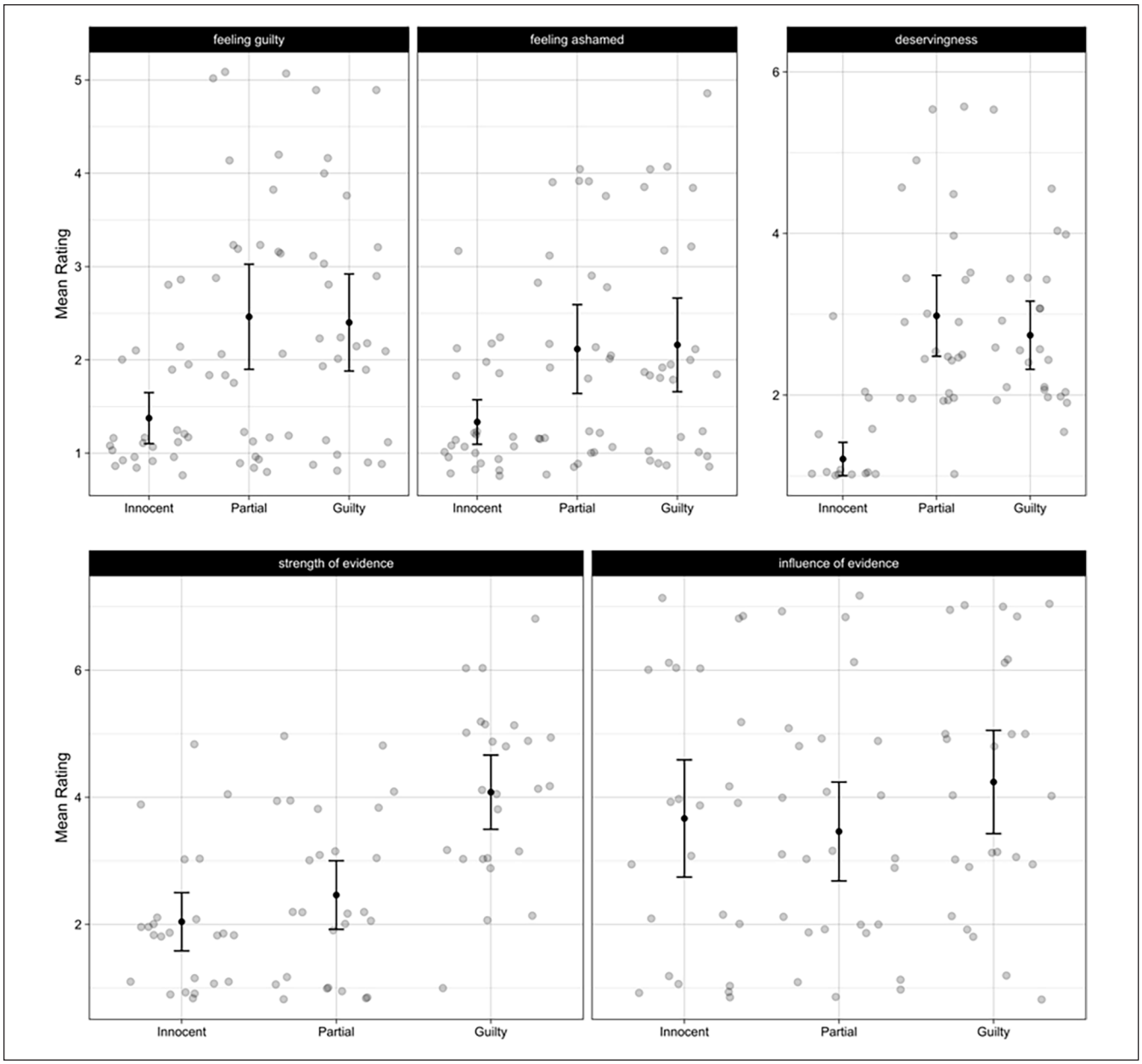

Figure 4. Distribution of Mean Ratings of Emotions and Evidence Judgments by Condition in Study 2.
Note. Dot plots: Error bars represent 95% CIs. Feelings of guilt and of shame measured using a single item each rated as part of the 20-item PANAS (1 = *very slightly or not at all* to 5 = *extremely*). Deservingness measured using two items (α = .89) in which labeled endpoints of Likert-type scales indicated that a rating of 1 was a low value (e.g., *not at all deserving*) and that a rating of 6 was a high value (*very deserving*). Labeled endpoints on Likert-type scales for evidence judgments ranged from 1 (*not at all*) to 7 (*a great deal*). CI = confidence interval; PANAS = Positive and Negative Affect Scale.

like innocent participants, feel wrongfully accused, and therefore less deserving, less guilty, angrier (*Hypothesis 5*; DeCelles et al., 2021), and plead guilty at lower rates than those accused of comparable transgressions. Second, we improved our measurement tools by using multi-item scales to measure guilt, shame, and anger, as well as judgments of deservingness of punishment and fairness of the plea deal and by administering these scales *prior* to making a plea decision. Third, based on Study 1, we constructed immersive role-playing stimulus materials designed to elicit a stronger emotional reaction in a less cognitively-taxing manner than a typical vignette study.

Study 3

Method

This study used a one-way (Guilty, Innocent, Partial-Comparable, Partial-Less Severe, Partial-More Severe) between-subjects design. An *a priori* power simulation using

the superpower package (Lakens & Caldwell, 2021) in *R* demonstrated that a total *N* of 750 participants would be sufficient to detect a small-sized effect of condition on guilty feelings (omnibus test of H_2; d_{cohen} =.175) with .99 power ($\alpha = .05$). The simulation using predicted means and SDs based on Study 2 and pre-testing is available on OSF (osf.io/8c79r).

Participants. We recruited 995 U.S., Canadian, and U.K. student participants through Prolific and Amazon's Mechanical Turk for a role-playing study. Participants were eligible adults who had a minimum 98% approval rate, were not a part of other crowd-sourcing platforms, had not participated in related pilot studies, and were paid US$2.75 (median survey duration = 21.48 min). Participants were excluded for failing to identify their condition (i.e., *stimulus attention check, n* = 83), failing an instructional attention check ($n = 168$), and failing a language proficiency check ($n = 5$). The final sample consisted of 746 participants (Innocent: $n = 157$; Guilty: $n = 156$; Partial-Comparable: $n = 154$; Partial-Accusation Less Severe: $n = 143$; Partial-Accusation More Severe: $n = 136$;): $M_{age} = 28.93$ years ($SD = 8.39$); 34.98% male, 61.53% female, 2.95% Other. Participants identified primarily as White (67.82%), followed by Asian (13.8%), Black/African American (5.63%), Hispanic/Latino (1.87%), and Other (10.8%). Some students (24.1%) reported having cheated in an academic context and a smaller amount reported having been accused of cheating (11.5%).

Measures

Emotional States. We used a slightly adapted version of the State Shame and Guilt Scale (SSGS-S: Marschall et al. 1994) to assess state feelings of guilt (6-items; e.g., "I feel remorse, regret") and shame (5-items; e.g., "I feel humiliated, disgraced") using a 5-point Likert-type scale ranging from (1 = *not feeling this way at all*, 3 = *feeling this way somewhat*, 5 = *feeling this way very strongly*). The average scores for the guilt ($\alpha = .95$ [.95, .96]) and shame ($\alpha = .90$ [.88, .91]) subscales were calculated. To measure anger, we used the 6-item Hostility subscale from the PANAS-X, which comprises self-report ratings of state emotions of hostility (e.g., "angry" and "irritable") uses a 5-point Likert-type scale (1 = *very slightly or not at all*, 5 = *extremely*) ($\alpha = .90$ [.89, .91]).

State Beliefs About Deservingness. To assess whether participants felt they deserved to be punished, we asked six questions adapted from Feather et al. (2011; $\alpha = .93$) such as "I deserve to be punished for my actions during the research session" (1 = *strongly disagree*, 5 = *strongly agree*; $\alpha = .97$ [.97, .98]).

Fairness. We assessed how fair participants thought the plea deal was relative to their alleged their actual misconduct (if in Guilty or PI condition), to the consequences of their alleged misconduct, and to the potential consequences of being found guilty by the Academic Integrity Board (1 = not at all fair, 4 = neutral, 7 = very fair). We computed a composite score using the average of the 3-items to which all participants responded $\alpha = .84$ [.82, .86]).

Procedure. After consenting and providing a nickname (which we then used to increase engagement with the materials), participants watched an immersive animated video (Minutes = 8' 29"-10') in which a narrator described the scenario to a gender-ambiguous avatar representing the participant. Voice actors representing Patty (Participant Partner), the Experimenter, and the Project Supervisor guided participant-avatars through events like those experienced by participants in Study 1[7] (see Figure 5). Participants were then randomly assigned to learn that their avatar had either followed the experimenter's instructions (*Innocent* condition) or that they had cheated on "Individual Task 2" (IT2) (*all other* conditions) (see Figure 6). Then, *Innocent* and *Guilty* participants were accused of cheating and stealing research compensation on IT2, whereas participants assigned to the *PI- Comparable Accusation* and the *More Severe Accusation* condition were accused of cheating and stealing research compensation on "Individual Task 1" (IT1)—a comparable offense. To make the accusation *more* severe, those in the *More Severe Accusation* group were also accused of stealing a research assistants' wallet containing her personal belongings and roughly $80 USD (see Figure 7). To make the accusation *less* severe, those in the *Accusation Less Severe* condition were accused of speaking with Patty, against the experimenter's rules, but not of cheating or stealing research compensation during IT1. After participants responded to three attention checks assessing their recall of the experimenter's instructions, the task they had cheated on (if any), and the task they had been accused of cheating on. Participants received immediate feedback on the accuracy of their response and were provided with the correct response if incorrect. Those who failed two or more attention checks were prevented from continuing the study.

Emotion Induction and Measurement. To induce an emotional response, participants responded to three open-ended questions about how they would feel and what they would think about (a) their behavior while completing the problem sets, (b) having to decide whether to take their case to the Academic Integrity Board, and (c) what they might say, if anything, to the project supervisor. The latter also served as an exploratory measure, not analyzed here, to capture spontaneous confessions. Participants then completed the SSGS-S, the Hostility Subscale, and the deservingness scale.

Plea Decision and Attributes of the Plea Judgments. Next, participants were presented with the same Study 1 plea offer (see Figure 8), decided to accept or reject the plea deal, were then asked to explain their choice and rated the perceived strength of the exculpatory and of the inculpatory evidence (1 = *not at all strong* to 7 = *very strong*), their estimated

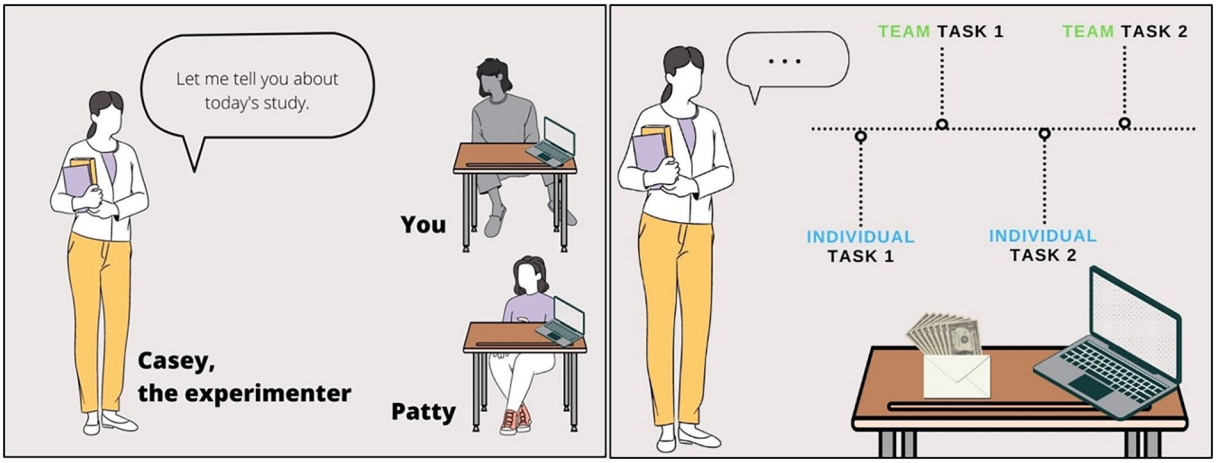

Figure 5. Images Depicting the Experimenter Introducing the Rules of the Study.
Note. All images were accompanied by audio of voice actors speaking directly to the participant-avatar. See the online article for the color version of this figure.

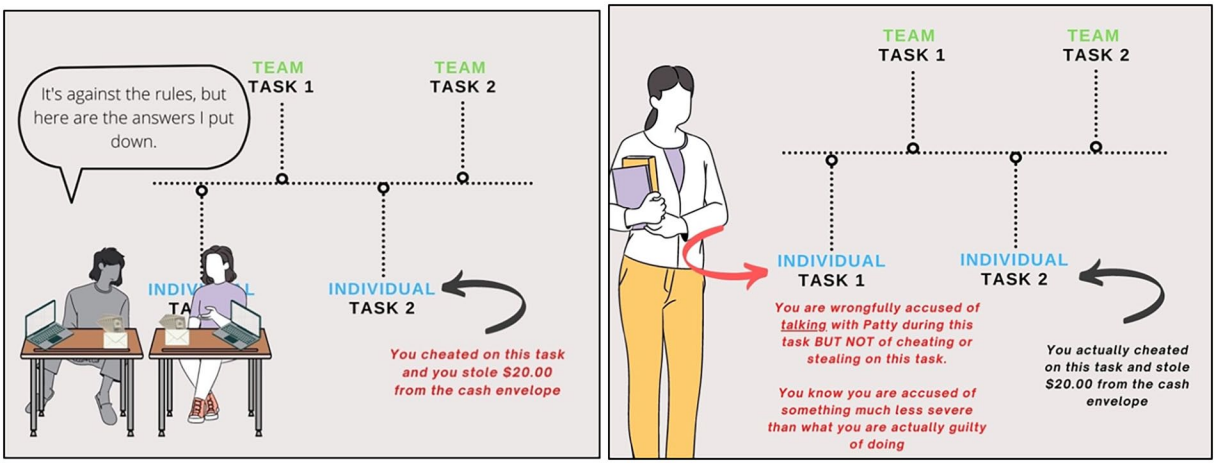

Figure 6. Images Depicting Guilty and Partially Innocent Participants Cheating With Patty (Left Image) and a Recap of the Accusation of Misconduct (Right Image).
Note. Innocent participant avatars see themselves obeying the instructors' participation rules instead (not pictured). Bottom panel depicts a recap of the participant-avatars' actual misconduct (Right side) and the experimenters' accusation for the Partially Innocent—Accusation Less Severe condition (Left side). Recaps such as these were presented *after* audio descriptions to avoid cognitively-taxing simultaneous displays of text and audio. See the online article for the color version of this figure.

likelihood of being found guilty (0%-100%), the severity of the misconduct they committed (if not innocent), and the severity of the misconduct of which they were accused (1 = *not at all serious* to 7 = *very serious*). Participants also rated the influence of three factors on their decision (1 = *not at all* to 7 = *very*)— the presence of inculpatory evidence, the consequences of accepting the deal (community service), the consequences of rejecting the deal (being referred to the Academic Integrity Board). Finally, they rated the fairness of the plea deal (4-items; 1 = *not at all fair* to 7 = *very fair*).

Demographics. Finally, participants indicated whether they had ever cheated (and if so, where they had been caught) in an academic context, their age, ethnicity, and gender, before being debriefed.

Results

Preliminary Analyses. Participants did not differ in age, gender, ethnicity, or past cheating by condition ($ps > .05$). Consistent with pilot testing, participants in the Less Severe Accusation

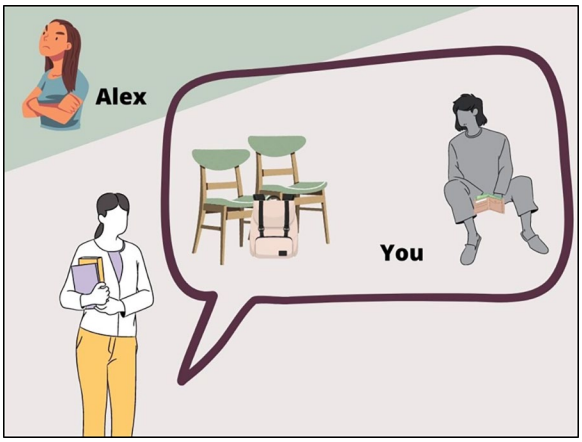

Figure 7. Image Depicting Partially Innocent Participants Accused of More Severe Misconduct by Stealing Alex's Wallet.
Note. See the online article for the color version of this figure.

group rated the severity of their accusation as *less* severe ($M = 3.17$, $SD = 1.65$) than those in the Comparable Accusation group, $M = 4.4$, $SD = 1.85$, $t(295) = 6.01$, $p < .001$, $d_{cohen} = -0.7$ 95% CI = [−0.96, −0.44], and those in the More Severe Accusation group ($M = 5.05$, $SD = 1.63$) rated their accusation as *more* severe than those in the Comparable Accusation group, $t(285) = 6.77$, $p < .001$, $d_{cohen} = 0.79$ [0.55, 1.05].[8] There were no significant differences in accusation severity ratings between the Innocent ($M = 4.64$, $SD = 1.6$), Guilty ($M = 4.38$, $SD = 1.73$), and PI-Comparable groups.

Confirmatory Analyses. Figure 9 shows the distribution and 95% confidence intervals of emotion, deservingness, and judgments of the plea ratings by culpability condition.

H_1: Culpability Predicts Plea Decision. Figure 1 shows the plea rates broken down by condition. There was a significant difference in plea decision across conditions $\chi^2(4) = 124.3$, $p < .001$. Consistent with Study 2, completely guilty participants accepted the plea at higher rates (73.71%) than those accused of committing a comparable, but different, transgression participants (59.74%) who in turn accepted the plea at higher rates than completely innocent (5.73%) participants. Further supporting H_1, participants accused of committing either a comparable or a a less severe transgression (57.34%) accepted the plea more often than those accused of comitting a more severe transgression (31.62%). See Table 1 for the logistic regression results.

H_2: Culpability Predicts Feelings of Guilt. On self-reported feelings of guilt, a one-way ANOVA revealed a significant effect of condition, $F(4, 741) = 112.37$, $p < .001$, $\eta^2 = .29$ [.24, .34]. Consistent with Study 1, feelings of guilt were significantly *greater* among Comparable Accusation participants compared with completely Innocent participants, $M_{diff} = 1.73$; $t(263) = 14.4$, $p < .001$, $d = 1.64$ [1.38, 1.90], and among Guilty compared with Innocent participants, $M_{diff} = 1.98$; $t(278) = 17.4$, $p < .001$, $d = 1.97$ [1.70, 2.24], but did not significantly differ between Comparable Accusation and Guilty participants, $M_{diff} = 0.25$; $t(306) = 1.82$, $p = .361$, $d = 0.21$ [−0.02, 0.43]. As expected, Less Severe Accusation participants did not report significantly *greater* feelings of guilt than Comparable Accusation participants, $M_{diff} = -0.12$; $t(295) = 0.91$, $p = .894$, $d = 0.10$ [−0.13, 0.33], or Guilty participants, $M_{diff} = -0.37$; $t(295) = 2.74$, $p = .051$, $d = 0.32$ [0.09, 0.55]. In support of H_2, participants in the More Severe Accusation condition feeling *less* guilty than those in the Comparable Accusation condition, $M_{diff} = -0.27$; $t(286) = 1.87$, $p = .335$, $d = 0.22$ [−0.01, 0.45], and significantly *more* guilty than Innocent participants, $M_{diff} = 1.46$; $t(234) = 12$, $p < .001$, $d = -1.57$ [−1.83, −1.31].

H_3: Culpability Predicts Feelings of Shame. On self-reported feelings of shame, a one-way ANOVA revealed a significant effect of condition, $F(4, 741) = 4.77$, $p < .001$, $\eta^2 = .03$ [.00, .05]. Similar to Study 1, feelings of shame did not significantly differ between Comparable Accusation participants and either Innocent ($M_{diff} = 0.31$; $t(308) = 2.35$, $p = .133$, $d = -0.26$ [−0.49, −0.04]) or Guilty participants ($M_{diff} = -0.19$; $t(308) = 1.35$, $p = 659$, $d = 0.16$ [−0.07, 0.38]). Again, Guilty participants reported feeling more ashamed than Innocent participants ($M_{diff} = 0.50$; $t(310) = 3.75$, $p = .002$, $d = -0.42$ [−0.65, −0.20]) and also, unexpectedly, more than Less Severe Accusation participants ($M_{diff} = 0.41$; $t(297) = 3.05$, $p = .021$, $d = -0.35$ [−0.58, −0.12]. Comparable Accusation participants did not significantly differ in reported feelings of shame from Less Severe Accusation ($M_{diff} = 0.22$; $t(295) = 1.66$, $p = 461.$, $d = -0.19$ [−0.42, 0.04]) or More Severe Accusation participants ($M_{diff} = -0.07$; $t(285) = 0.51$, $p = .987$, $d = 0.06$ [−0.17, 0.29]).

H_4: Culpability Predicts Feelings of Deservingness of Punishment. Next, an ANOVA indicated that feelings of deservingness differed across condition, $F(4, 741) = 464.31$, $p < .001$, $\eta^2 = .51$ [.46, .55]. Results were consistent with H_4: Participants accused of a comparable transgression did not feel significantly more deserving of punishment than those accused of a *less* severe transgression, $M_{diff} = 0.15$; $t(290) = 1.12$, $p = .797$, $d = -0.13$ [−0.35, 0.10], or than those who were completely guilty of the transgression of which they were accused, $M_{diff} = -0.33$; $t(308) = 2.66$, $p = .062$, $d = 0.31$ [0.08, 0.53], who felt the most deserving. Despite engaging in similar misconduct, participants accused of a more severe transgression felt less deserving of punishment than those accused of committing a comparable transgression, $M_{diff} = -0.59$; $t(279) = 4.54$, $p < .001$, $d = 0.53$ [0.30, 0.77], a less severe transgression, $M_{diff} = 0.45$; $t(276) = 3.30$, $p < .001$, $d = 0.39$ [0.16, 0.63], and than those who were completely guilty of the accused transgression, $M_{diff} = -0.92$; $t(278) = 7.09$, $p < .001$, $d = 0.83$ [0.60, 1.07].

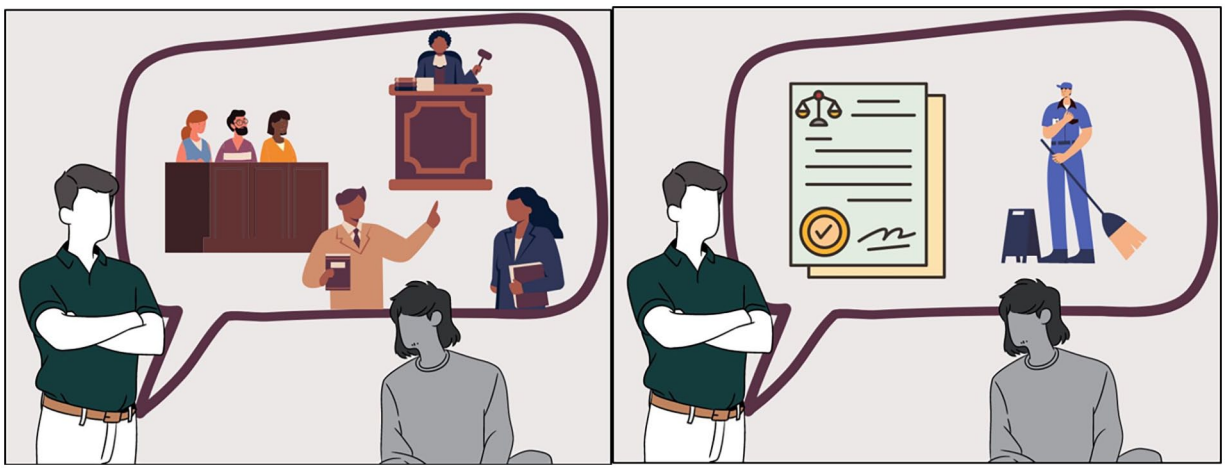

Figure 8. Project Supervisor Informs Describes the Academic Integrity Board (Left Image) and the Plea Deal (Right Image).
Note. See the online article for the color version of this figure.

As expected, Innocent participants felt the least deserving, even compared with participants in the More Severe Accusation condition who had also been wrongfully accused, $M_{diff} = -1.86$; $t(165) = 18$, $p < .001$, $d = 0.94$ [1.93, 2.51].

H_5: Culpability Predicts Feelings of Anger. An ANOVA indicated that feelings of anger were significantly different across condition, $F(4, 741) = 16.13$, $p < .001$, $\eta^2 = .08$ [.05, .12]. Results were largely consistent with H_5: In line with our predictions, reported feelings of anger did not significantly differ between completely innocent participants and participants accused of a more severe transgression , $M_{diff} = -0.16$; $t(287) = 1.29$, $p = .699$, $d = 0.15$ [−0.08, 0.38]. Also, as expected participants accused of committing a less severe transgression were significantly less angry than those accused of a more severe transgression, $M_{diff} = -0.54$; $t(266) = 4.61$, $p < .001$, $d = 0.55$ [0.31, 0.79], but, contrary to H_5, not less so that participants accused of a comparable transgression, $M_{diff} = -0.03$; $t(295) = 0.29$, $p = .999$, $d = 0.03$ [−0.20, 0.26], or than guilty participants, $M_{diff} = -0.03$; $t(297) = 0.27$, $p = .999$, $d = 0.03$ [−0.20, 0.26].

H_6: Planned Mediation Paths to Guilty Pleas. We performed a pre-registered multiple mediation analysis (similar to the one tested in Study 1) to examine two emotional routes—guilt and shame—through which culpability could result in a guilty plea. We used dummy-coded culpability contrasts, comparing the No Cheat condition (Reference Group: Innocent) against each of the Cheat conditions: *Innocent versus PI-C, Innocent versus Guilty, Innocent versus PI-LSA, Innocent versus PI-MSA*. Results were consistent with our predictions (H_6) and the pattern of results from Study 1. Increased feelings of *guilt* (b_1 path: $b = .12$ [.08, .16], $SE = .02$, $p < .001$) predicted plea decision, whereas *shame* did not (b_2 path: $b = -.02$ [-.06, .02], $SE = .02$, $p = .367$), even when controlling for culpability. Feelings of *guilt* mediated the relationship between each culpability contrast and plea decision ($ps < .001$), whereas feelings of *shame* did not ($p > .398$).

Specifically, whereas the indirect effect of feelings of *guilt* was significant in the *Innocent versus PI-C* (X1) contrast (a_1*b_1: $b = 0.20$ [0.13, 0.29], $p < .001$), the indirect effect through feelings of *shame* was not significant (a_5*b_2: $b = -0.01$ [−0.02, 0.01], $p = .432$). A comparison of the two relative indirect effects through guilt and shame demonstrated that these were significantly different, $b = 0.21$ [0.13, 0.31], $p < .001$. Similarly, in the *Innocent versus Guilty* (X2) contrast the indirect effect through feelings of guilt was also significant (a_2*b_1: $b = 0.24$ [0.15, 0.33], $p < .001$), whereas the indirect effect through feelings of *shame* was not (a_6*b_2: $b = -0.01$ [−0.03, 0.01], $p = .398$). Here again, a comparison of the two relative indirect effects demonstrated that these were significantly different ($b = 0.25$ [0.14, 0.35], $p < .001$). Results across other culpability contrasts were highly consistent: the comparison of the relative indirect effects of shame and guilt was significant in the *Innocent versus PI-LSA* (X3), $b = 0.19$ [0.12, 0.28], $p < .001$, and the *Innocent versus PI-MSA* (X4) contrast, $b = 0.18$ [0.10, 0.27], $p < .001$. Figure 10 shows the results of this model. Supplemental Figure 2 shows the same model with PI-C as the Reference Group. Unsurprisingly, this latter model demonstrates that the relative indirect effect of guilt did not significantly differ in any contrast of *PI-C versus Other Cheat Condition (Guilty, PI-LSA, or PI-MSA), ps* > .078 - .039, as guilty feelings did not differ between these groups rendering the paths non-significant.

These findings suggest that compared with being entirely innocent, being responsible for some wrongdoing—even when that wrongdoing was *less* severe, merely comparable, or *more* severe than the target of the accusation—is linked to increased feelings of guilt and shame; and the more guilty,

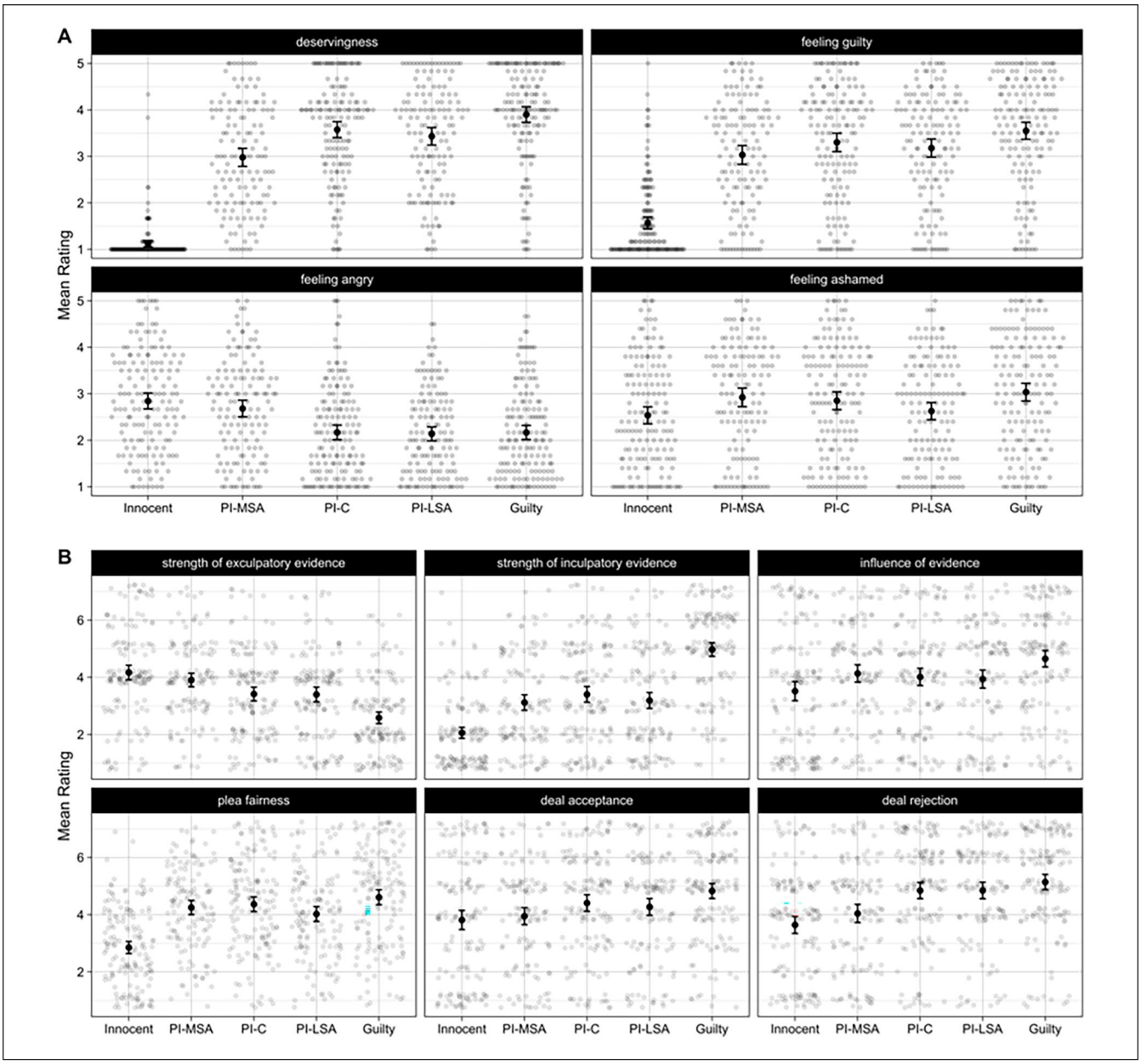

Figure 9. Distribution of Mean Ratings of Emotions and Judgments of Plea Attributes by Condition in Study 3.
Note. (A) Bee swarm plot shows individual mean ratings on emotion (1 = *not feeling this way at all*, 3 = *feeling this way somewhat*, 5 = *feeling this way very strongly*) and dessert scales (1 = *strongly disagree* to 5 = *strongly agree*). (B) Jittered plot of single-item ratings (except for plea fairness scale). Labeled endpoints of Likert-type scales indicated that a rating of 1 was a low value (e.g., *not at all*) and that a rating of 7 was a high value (e.g., *very*). Higher density of ratings is represented by thicker areas of the plot. Error bars in (A) and (B) depict 95% CIs. Supplemental Table 5 contains the ANOVA results for these measures. PI-MSA = Partially Innocent—More Severe Accusation; PI.-C = Partially Innocent—Comparable Accusation; PI—LSA = Partially Innocent—Less Severe Accusation; CI = confidence interval.

but not ashamed, people felt, the more willing they were to accept the plea offer.

Exploratory Serial Mediation. Next, we tested an exploratory serial mediation model to assess whether the effect of culpability on plea decision is serially mediated by judgments about deservingness of punishment and feelings of guilt.[9] We based our mediation model on Feather's (1999) theory of deservingness. According to this theory, an individual's evaluation of deservingness and the emotions associated with it depend on the consistency between their actions and the resulting outcomes. When a positive action leads to a positive outcome, it is judged as deserved, resulting in feelings of pride and pleasure. Similarly, a negative action that leads to a negative outcome is also judged as deserved but results instead in feelings of guilt and regret. Feather et al. (2011)

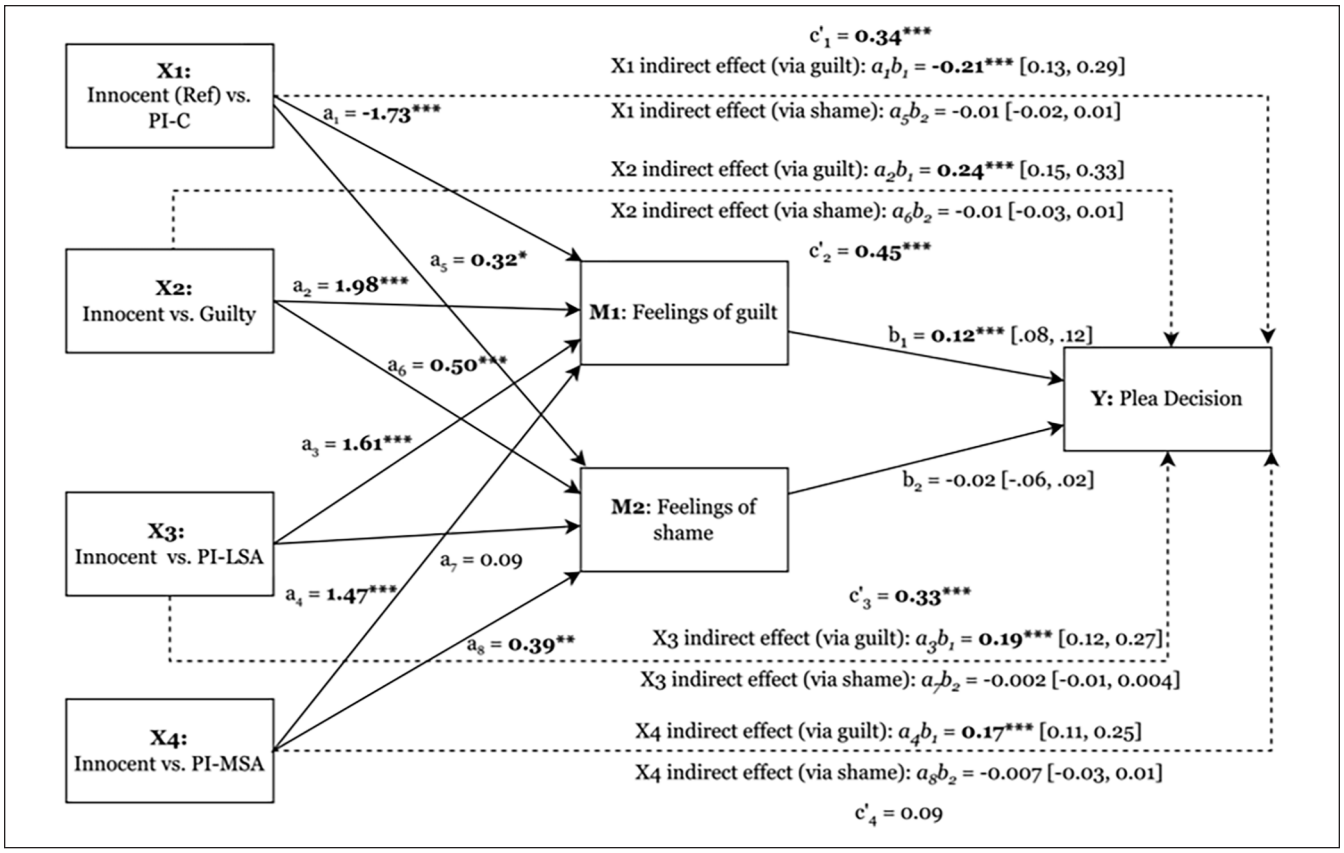

Figure 10. Study 3 Parallel Mediation Model of Culpability Predicting Plea Decision Mediated by Feelings of Guilt, But Not Shame.
Note. Results of a parallel mediation analysis examining the relative indirect effects of condition on dummy-coded culpability contrasts (X1, X2, X3, and X4) on plea decision (Y) through state feelings of guilt (M1) and shame (M2) composite scales. Unstandardized estimates are displayed, with 95% confidence intervals; c' = direct effect of X contrast on Y. Reference Group: Innocent: = 1, All others = 0; Plea Decision: Accept = 1, Reject = 0. Estimated coefficients are based on bootstrapping procedure with 10,000 bootstrap samples. n = 746. PI-C = Partially Innocent (Comparable Accusation); PI-LSA = Partially Innocent (Accusation Less Severe); PI-MSA = Partially Innocent (Accusation More Severe).
*$p < .05$. **$p < .01$. ***$p < .001$

found support for this model by demonstrating that participants' appraisals of deservingness mediated the relationship between negative actions resulting in negative outcomes, and emotions such as guilt, regret, resentment, and anger. Indeed, in the current work, participants' beliefs about deservingness were strongly correlated with feelings of guilt, $r(746) = .64$ [.60, .68], $p < .001$. The current analysis expanded Feather et al.'s (2011) deservingness model by incorporating a willingness to take responsibility measure.

To expand on this model, we investigated whether the serial indirect effect of each of four culpability contrast (X1, X2, X3, X4) on plea decision (Y) through deservingness (M1) and then guilt (M2) would be significant. To compute the serial indirect effect, we estimated three logistic regression models. In the first model, we regressed the first mediator (dessert) on the four culpability contrasts. In the second model, we regressed the second mediator (feelings of guilt) on the four culpability contrasts and the first mediator. In the third model, we regressed the dependent variable (plea decision) on the four culpability contrasts and the two mediators. Figure 11 shows the results of this model.

Consistent with H_2 and H_4, in the first and second mediator models, all coefficients were positive and significant indicating that in each of the Cheating groups participants believed themselves more deserving of punishment and reported feeling more guilty than participants in the Innocent group ($p < .001$). Beliefs about deservingness were positively associated with feeling guilty ($b = 0.50$ [0.42, 0.58], $p < .001$), even controlling for condition. In the third model, the effect of deservingness beliefs ($b = 0.18$ [0.14, 0.21], $p < .001$) and feelings of guilt ($b = 0.04$ [0.01, 0.07], $p = .018$) on plea decision were each significant. In addition, whereas the *Innocent versus Guilty* and *Innocent versus PI-MSA* contrasts were significant, the *Innocent versus PI-C* and *Innocent versus PI-LSA* contrasts were not significant. The latter results may tentatively suggest that the mediators fully explain the variance in plea decision between *Innocent versus PI-C* and between *Innocent versus PI-LSA* conditions.

Next, we computed the coefficient for the relative serial indirect effect of each culpability contrast through deservingness beliefs and guilty feelings as the product of these coefficients (e.g., Innocent vs. Cheat Condition → Deservingness,

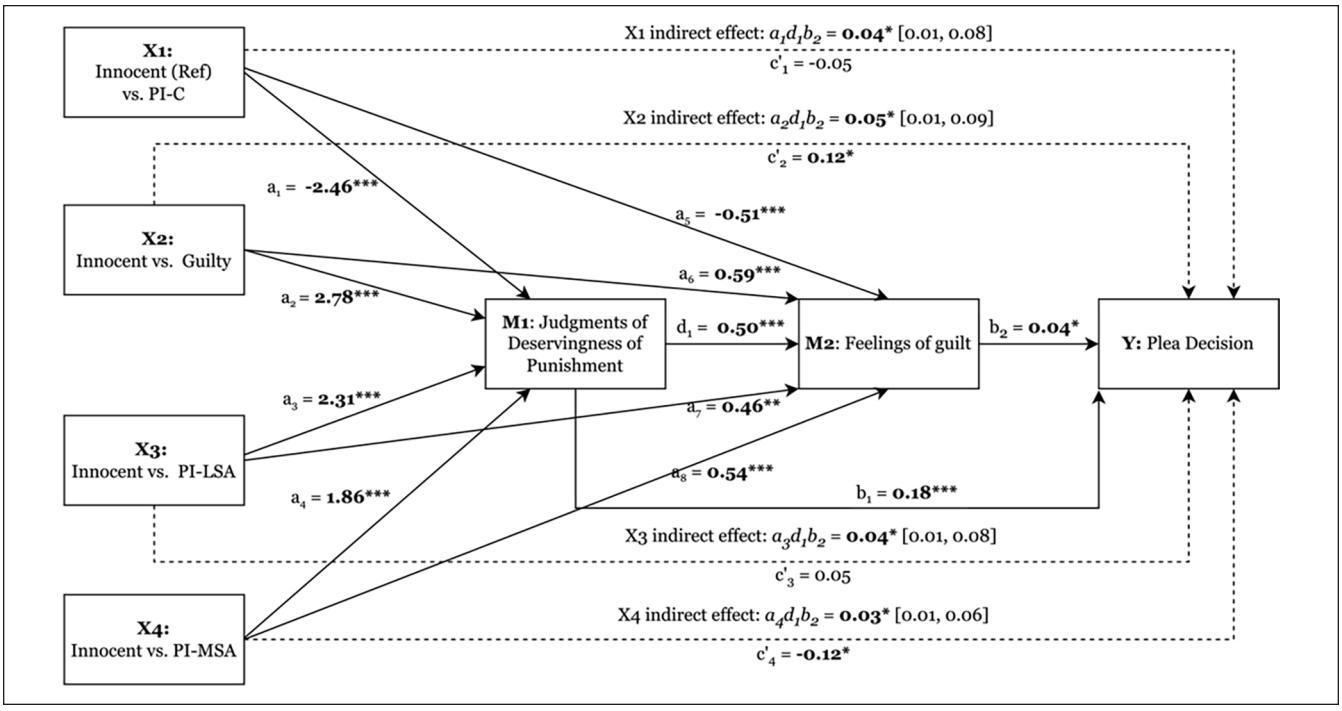

Figure 11. Study 3 Serial Mediation Model of Culpability Predicting Plea Decision Mediated by Judgments of Deservingness of Punishment and Feelings of Guilt.

Note. Results of serial mediation analysis examining the relative indirect effect of condition contrasts (X1, X2, X3, and X4) on plea decision (Y) through judgments of deservingness of punishment (M1) and state feelings of guilt (M2). Unstandardized estimates are displayed, with 95% confidence intervals. Reference Group: Innocent: = 1, All others = 0; Plea Decision: Accept = 1, Reject = 0. Estimated coefficients are based on bootstrapping procedure with 10,000 bootstrap samples. n = 746. PI-C = Partially Innocent (Comparable Accusation); PI-LSA = Partially Innocent (Accusation Less Severe); PI-MSA = Partially Innocent (Accusation More Severe).
*$p < .05$. **$p < .01$. ***$p < .001$.

Deservingness → Guilty Feelings, Guilty Feelings → Plea Decision). We then used a bootstrapping procedure implemented with 10,000 samples in *lavaan* to assess whether these relative indirect effects were significant. The indirect effects of all culpability contrast on plea decision through deservingness beliefs and feelings of guilt were significant. Supplemental Figure 3 shows the same model with PI-C as the Reference Group.

In sum, these findings suggest that when accused of wrongdoing they did *not* commit, participants who imagined having committed *another* wrongdoing believed themselves to be more deserving of punishment relative to completely innocent individuals, and that beliefs about deservingness of punishment were, in turn, associated with feeling guilty, which in turn was associated with a greater willingness to plead guilty.

Exploratory Judgments of Plea Attributes. Figures 9B and 12 show the distribution of mean ratings for judgments of the plea attributes by culpability condition. Visual inspection of mean ratings suggests that PI groups provided more middling-range responses relative to completely Innocent or Guilty participants: for example, compared with Innocent participants, PI groups rated the inculpatory evidence as *stronger*, the exculpatory evidence as *weaker*, and estimated their likelihood of conviction as *higher*. Whereas compared with Guilty participants, PI participants rated the inculpatory evidence as *weaker*, the exculpatory evidence as *stronger*, and estimated their likelihood of conviction as *lower*. For ANOVA results, see Supplemental Table 5. Bivariate correlations between judgments of the plea experience and plea decision split by condition are available on OSF (Guilty: osf.io/3d8cr; Innocent: osf.io/zmqts; PI-Comparable Accusation: osf.io/juf9t; PI-Less Severe Accusation: osf.io/y4rg2; PI-More Severe Accusation: osf.io/j4zmr).

Discussion of Study 3

The results of Study 3 replicated and extended those of Study 1 using more valid multi-item measurement tools. Consistent with our pre-registered hypotheses, PI participants accused of comparable or *less* severe wrongdoing compared with their actual misconduct reported beliefs about deserving punishment, and, in turn, feelings of guilt that were not significantly different from those of guilty participants, but greater than those of innocent participants. Greater feelings

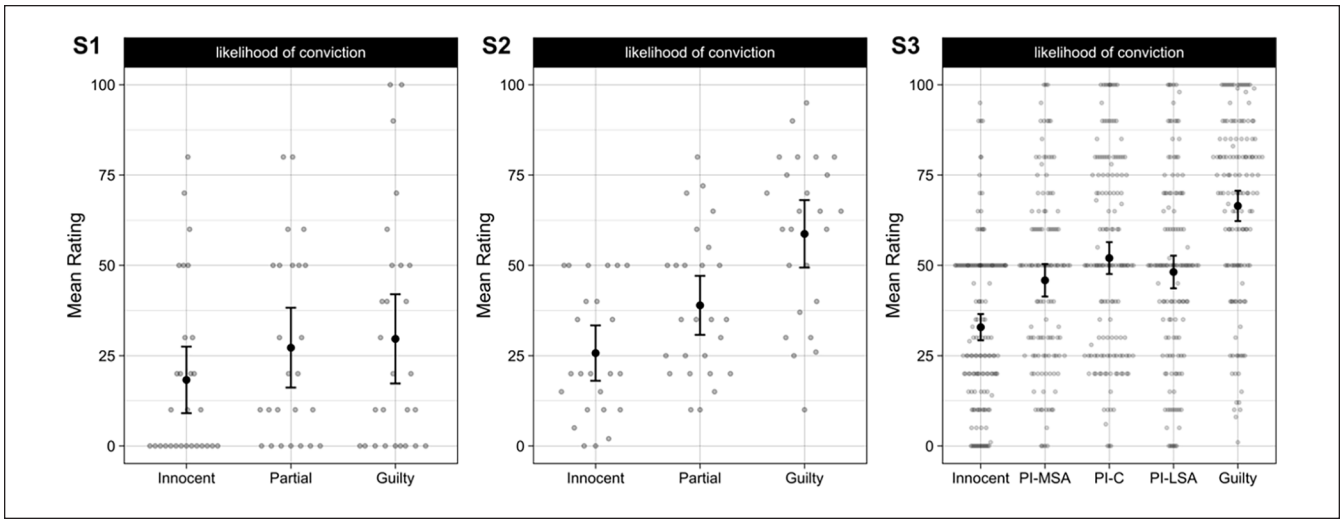

Figure 12. Distribution of Estimates of Likelihood of Conviction Across All Studies.
Note. Dot plot: Error bars represent 95% CIs. Labeled endpoints of Likert-type scales indicate that an estimate of 0% chance indicated a "I would definitely be acquitted," a 50% chance indicated "unsure," and a 100% chance indicated "I would definitely be found guilty." S1 = Study 1; S2 = Study 2; S3 = Study 3; CI = confidence interval; PI-C = Partially Innocent—Comparable Accusation; PI-LSA = Partially Innocent—Less Severe Accusation; PI-MSA = Partially Innocent—More Severe Accusation.

of guilt—but not of shame—were, in turn, associated with increased plea acceptances. In addition, PI participants accused of wrongdoing *more* severe than their actual misconduct reported feeling as angry as innocent participants and accepted the plea significantly less often than those accused of comparable wrongdoing. But those accused of *less* severe wrongdoing accepted the plea nearly as often as those accused of comparable misconduct.

General Discussion

The goals of the current study were to investigate the phenomenology and decision-making of partially innocent individuals. Three experiments demonstrate that PI participants—in sharp contrast to those who were completely innocent— often pleaded guilty to an offense they did not commit and did so at comparable or slightly lower rates as participants who were in fact guilty of that offense. Consistent with Zottoli et al.'s (2016) field study, PI participants pled guilty despite the high-stakes negative consequences (Study 1) and despite being accused of a more severe or a less severe offense (Study 3). On the secondary goal of investigating the psychology of PI defendants' plea decisions, we found support for our prediction that PI individuals would feel more deserving of punishment (Studies 2 and 3) and experience greater feelings of guilt than innocent individuals. Feeling guilty, in turn, was associated with a higher likelihood of accepting the plea deal even when controlling for state feelings of shame (Studies 1 and 3). Whereas beliefs about just deserts and guilty feelings were closely associated with plea decisions, dispositional beliefs about deserving negative outcomes and personality traits, like interrogative compliance, just world beliefs, and guilt and shame proneness, were not (Study 1).

Even so, PI participants differed from guilty participants in three significant ways. First, whereas guilty participants reported being more culpable than innocent participants, PI participants rated their own culpability status as somewhere between 'completely guilty' and 'completely innocent'. Second, when asked to report which task they had cheated on, PI participants denied cheating far more often than guilty participants—by a margin of 20.7% to 58.3%. Third, compared with guilty participants, PI participants regarded the evidence in their favor as *stronger* and the evidence against them as *weaker*, ultimately estimating a *lower* likelihood of conviction (Studies 2 and 3). PI individuals seemingly calibrated their evaluation of the strength and influence of the evidence on their plea decisions in light of their own private knowledge of their partial culpability. This is important because evidence strength—via its influence on estimates of likelihood of being found guilty at trial—is a predominant factor in defendant decision-making (e.g., Kramer et al., 2007; Redlich et al., 2016).

In summary, this study provides strong preliminary support for the notion that, at least behaviorally and emotionally, PI individuals may resemble guilty individuals more than completely innocent individuals. Notably, however, PI people maintained a private understanding of their own ambivalent state of culpability (particularly when accused of committing a worse transgression), where the opposition's evidence might fall short, and why ambiguity might benefit them in a final adjudication.

Strengths, Limitations, and Future Directions

This work is the first to empirically demonstrate the effect of partial innocence on responsibility-taking in both real and imagined plea contexts. However, there are limitations that bear discussing. For one, the elevated ratings of guilt and shame as well as increased plea acceptances across conditions in the role-playing studies suggest it may have been somewhat difficult for student–participants to imagine how they would feel and react in response to a genuine accusation of academic misconduct. Indeed, research consistently finds that people tend to overestimate the impact of events on their emotional states (see Ayton et al., 2007; Kawakami et al., 2009; Wilson & Gilbert, 2013). However, this same body of work consistently finds that people can, at minimum, accurately predict the emotional valence resulting from events, as well as identify which events will have a greater impact on their emotions. In fact, some work suggests that results obtained from studies that use hypothetical scenarios align with those from studies that examine real emotional experiences (Robinson & Clore, 2001). Considering that academic misconduct is not entirely uncommon among students (25% had previously cheated in Study 3), it is likely that predictions about the direction, if not the magnitude, of expected feelings and behaviors might be at least partially accurate. In support, the relative differences in feelings of guilt and plea decisions across conditions were consistent across three studies using varied methodologies, yielding large effect sizes.

Furthermore, in line with Ajzen's (1991) theory of planned behavior, prior work suggests that intentions to engage in a behavior can accurately predict actual behavior when certain conditions—all met in the current work—are present, such as the availability of variability in responses, compatibility between the intention and behavioral measure, and controllability of one's own behavior (Ajzen, 2020, pp. 320–321; Sheeran, 2002). Still, this work finds that people's intentions tend to become biased toward socially desirable responses when a behavior (such as behaving ethically and taking responsibility for one's wrongdoing) is deemed more socially normative (Sheeran, 2002). This social desirability bias may also explain the higher reported feelings of guilt and responsibility-taking in the curent role-playing compared with real-world decision-making studies.

Beyond describing the phenomenology of partial innocence, this work's primary theoretical contribution lies in its consistent support of a novel moral emotional explanation of plea decision-making. Study 3 operationalized "partial innocence" broadly by accusing participants of transgressions that were either ethically comparable, *less* severe or *more* severe than their actual misconduct. This enabled us to assess whether the proposed emotional mechanism could explain the self-incriminatory plea decisions of various types of PI persons. As expected, though the consequences for pleading guilty were held constant, PI persons who were wrongfully accused of *more* severe wrongdoing than their actual misconduct, were nearly as angry as completely innocent participants and significantly angrier and less willing to plead guilty than those accused of engaging in comparable misconduct. The fact that more severe wrongful accusations caused otherwise guilty persons to enter this partial innocence state where they felt less deserving of punishment, less guilty, angrier, and less willing to take responsibility reveal an important boundary condition to the effects found here.

Indeed, this finding highlights the value of gaining insight into the legal experience of being overcharged. Overcharging occurs when prosecutors threaten defendants with excessively severe charges that carry disproportionately harsher punishment than is warranted by the defendants' alleged misconduct to dissuade them from exercising their right to trial (Caldwell, 2011). These findings may suggest that when making accusations, a failure to map the magnitude of the charges to the accusation could backfire, causing the accused to no longer desire to take responsibility for their actual wrongdoings. In the real world, increasing the severity of charges is typically associated with increased punishment which may also deter responsibility-taking. Conversely, other work finds that overcharging by threatening guilty mock defendants and their mock attorneys with excessively harsh (compared with more appropriate) sentences increases plea acceptance (Cardenas, 2023; c.f. Schneider & Zottoli, 2019). Given these apparent discrepancies, future work should investigate the unique and interactive effects of culpability and type of overcharging (qualitatively versus numerically severe) on responsibility-taking.

Although most past research has examined defendant plea decisions using exclusively hypothetical vignettes or field study self-reports (c.f. Cardenas, 2023; Dervan & Edkins, 2013; Henderson & Levett, 2018; Pardieck et al., 2020; Perillo et al., 2014; Wilford, Sutherland, Gonzales, & Rabinovich, 2021; Wilford, Wells, & Frazier, 2021; Wilford & Wells, 2018), the current work used a triangulation approach that combined a high-stakes real-world decision-making laboratory paradigm that afforded greater experimental control and then replicated and extended findings in two immersive and interactive role-playing studies. Where results differed was in the more extreme responses of participants in the role-playing compared with real-world experimental paradigms, who reported a lower likelihood of conviction, feeling less guilt, and ultimately pleaded guilty less often. The more pessimistic estimates of role-playing participants are consistent with research on the *illusion of transparency*, the belief that one's internal states, thoughts, and emotions are visible to others (Gilovich et al., 1998). Indeed, given that the methodology instructed them to reflect on their thoughts and emotions, role-playing participants may have implicitly believed that the truth of their transgressions would be evident to an external disciplinary committee. In contrast, the more optimistic judgments of participants making ostensibly real-world decisions are more consistent

with the known tendency of individuals to make self-serving evaluations under conditions of uncertainty, for example by viewing themselves more positively (Dunning et al., 1989; Epley & Dunning, 2000), behaving as if innocent when their culpability is ambiguous (Tor et al., 2010, Studies 4 and 5), and by denying case facts, their knowledge, culpability, or the amount of harm caused (see Bibas, 2004). Discrepant findings can cause concerns over the reliability of the findings. Yet despite vast methodological differences, including the use of different accusations, measurement instruments, samples, consequentiality, format and more, results related to the moral emotional pathways to guilty pleas and relative differences between groups were highly consistent across studies. This consistency suggests that certain simulation studies may closely model the emotions, cognitions, and behaviors of student-participants involved in more expensive, time-costly, high-stakes deceptive plea decision-making paradigms. To further investigate the utility of this triangulation approach for understanding plea decision-making, future work should continue to compare results from behavioral observations of situations requiring responsibility-taking against results from analogous hypothetical scenarios (e.g., Baumert et al., 2013).

Understanding how different types of partially innocent individuals think, feel, and behave in legal contexts is important as numerous structurally embedded charging practices, such as strategic overcharging, can give rise to partial innocence. Yet future research could also examine whether the current explanatory model explains the phenomenology of real and imagined partial innocence wrought from nonstructural factors. For example, a PI defendant may be one who was ignorant of the law, or one who knowingly violated a law they judged to be unjust and unfair; a defendant may lack clarity over their culpability during the time of a crime (resulting from intoxication or some other compromised mental state); or judge themselves to be morally guilty, and therefore worthy of, punishment as in sudden infant death cases which have contributed to false confessions and the wrongful conviction of numerous bereaved mothers such as Jacqueline Fletcher who falsely confessed under pressure that she drowned her infant son because she felt responsible for his well-being (Gudjonsson, 2003). The current work should examine whether people like Fletcher might be especially vulnerable to the confrontation-based Reid interrogation technique commonly used in the United States and Canada (Davis & O'Donohue, 2004; Inbau et al., 2013). This interrogation tactic instructs law enforcement to exploit the "troubled conscience. . . and. . . moral guilt" of "emotional offenders" by introducing minimization themes that emphasize sympathy and understanding (p. 185).

Given the foundational nature of this work, one might reasonably wonder if there truly does exist a singular "phenomenology of partial innocence," or if the experience of partial innocence might systematically vary depending on factors such as the circumstances that resulted in one being partially innocent.[10] Perhaps like other well-studied psychological phenomena (e.g., anchoring biases, Turner & Schley, 2016; cognitive biases, Oeberst & Imhoff, 2023), partial innocence may encompass various phenomenological experiences characterized by similar outcomes (i.e., greater responsibility-taking relative to innocence). Future work should investigate whether other forms of partial innocence that result in different phenomenological experiences are the result of different non-mutually exclusive psychological causes (e.g., emotion-based, biased information processing).

Conclusion

This work is the first to suggest that PI individuals guilty of committing a transgression may take responsibility for a transgression *not* committed to assuage feelings of guilt. Given that the proportion of PI defendants potentially dwarfs that of innocent defendants, it is critical that we continue to advance our understanding of the phenomenology of this group at various stages of their involvement with the criminal legal system.

Authors' Note

Portions of this study were presented at the 2019 annual meeting of the European Association for Psychology and Law in Santiago de Compostela, Spain, the 2020 annual meeting of the American Psychology and Law Society, and of the 2020 Justice and Morality Preconference of the Society for Personality and Social Psychology in New Orleans, LA, USA.

Acknowledgments

We would also like to thank our undergraduate research assistants—Lucrezia Rizzeli, Angelo Luongo, Devon Kaat, Eric Korzun, Elena Christofi, Danielle Strolia, Athena Sher, Helen Gavrilov, Brooke Flagler, Rosalba Linares, Emma Draper, and Devika Goel—for their invaluable help with data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this research was obtained from the National Science Foundation, Williams College DRFC, and the American Psychology-Law Society awarded to the first author, as well as a John Jay College PSC-CUNY Grant awarded to the first and third authors. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID iDs

Stephanie A. Cardenas https://orcid.org/0000-0002-1148-2573
Patricia Y. Sanchez https://orcid.org/0000-0002-4464-8365

Supplemental Material

Supplemental material is available online with this article.

Notes

1. The choice to refer to these individuals as partially *innocent* rather than partially *guilty* is in line with the U.S. legal system's presumption of innocence until proven guilty. Further, though such individuals may consider themselves (or be perceived by others as) morally or ethically guilty of wrongdoing, because they are as a matter of legal fact innocent of the charges of which they have been accused, we refer to them as partially innocent.
2. Ten trained research assistants and the first and second authors (three female experimenters, four female confederate partners, three male project supervisors) helped run sessions. RAs were trained by viewing recordings of pilot sessions and conducting mock sessions with each other under the supervision and guidance of the first and second authors. Pre-set randomizers were used to randomly assign experiments to a type of accusation (Individual Task 1 vs. 2: see osf.io/js5tw), as well as confederate-partners to a cheating inducement (cheat vs. no cheat: osf.io/er7yf). All personnel were randomly assigned to sessions based on scheduling availability. However, to ensure that confederate behaviors were consistent across conditions, all interactions between participants and research personnel were scripted (for protocol and script, see osf.io/hjnw8). Participant evaluations of their partners and the experimenter provide indirect evidence of perceived consistency across research personnel (for information about these scales, see osf.io/uhq25). Specifically, participant ratings of the experimenter did not significantly differ based on experimenter, $F(2, 82) = 1.42$, $p = .246$, $\eta^2 = 0.04$. Similarly, partner ratings did not differ based on partner, $F(3, 81) = 0.56$, $p = .638$, $\eta^2 = 0.02$. Finally, there was no association between any partner (p =.07), experimenter (p =.06), or project supervisor (p =.93) and participant plea decision.
3. To ascribe any effects on plea decision to feelings of guilt and shame rather than negative affect more generally, we compared a composite score of general negative affect (e.g., fearful, upset, nervous) after plea decision and found no differences across conditions, $F(2, 79) = 0.11$, $p = .89$.
4. Feelings of guilt, $b = 0.77$, $z = 11.4$, $p < .001$, $OR = 2.16$ [1.90, 2.48], and of shame, $b = 0.38$, $z = 5.8$, $p < .001$, $OR = 1.46$ [1.28, 1.65], were independently strongly associated with plea decision.
5. The advantage of this analytic plan when compared with the parallel mediation procedure recommended by Hayes and Preacher (2014) is that it allowed us to retain the dichotomous dependent variable as Ordinary Least Squares regressions employed by the SPSS PROCESS Macro require continuous dependent variables. We tested the statistical significance of the relative indirect and direct effects using the 95% bootstrap confidence intervals. Relative here refers to the total, direct, and indirect effects in single and multiple mediation models containing multicategorical independent variables because it calculates the effect of being in one group relative to the reference group (Innocence).
6. The H_2 and H_3 effect sizes correspond to the difference in feelings of guilt between Partially Innocent and Innocent participants and between those who accepted and those who rejected the plea, respectively.
7. Videos for each condition are available on the OSF: Guilty (osf.io/k3bcg); PI-Comparable (osf.io/y5eua); PI-More Severe Condition (osf.io/uwy3m); PI-Less Severe (osf.io/dkpb7); Innocent (osf.io/r8tpq).
8. Pre-testing ($n = 64$) demonstrated that participants regarded the Less Severe Accusation as less serious, less consequential for the researchers, and less worthy of punishment ($M_{composite} = 3.9$ of 7, $SD = 1.69$) compared with the Comparably Severe Accusation, $M_{composite} = 5$, $SD = 1.18$, $t(63) = -5.63$, $p < .001$, $d = 0.76$ [0.40, 1.11]. As expected, participants also regarded the More Severe Accusation as more serious, more consequential, and more worthy of punishment ($M_{composite} = 5.76$, $SD = 0.97$) compared with the Comparably Severe Accusation, $t(63) = -5.41$, $p < .001$, $d = -0.70$ [−1.06, −0.35].
9. We pre-registered that if beliefs about deservingness were independently associated with plea decision, we would add desert as a third indirect pathway to the exploratory serial mediation model reported above. A logistic regression was significant—$\chi^2(1) = 200.9$, $p < .001$, $R^2 = .46$—and indicated that deservingness is strongly independently associated with plea decision, $b = 1.15$, $z = 14.17$, $p < .001$, $OR = 3.16$ [2.71, 3.72]. However, we report a serial mediation here instead as conceptually this model is more in line with prior theorizing and empirical work on the causal role of desert on affective responses to outcomes (e.g., Feather, 1999; Feather et al., 2011).
10. We thank an anonymous reviewer for suggesting this point.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. *Human Behavior and Emerging Technologies*, *2*(4), 314–324. https://doi.org/10.1002/hbe2.195

Ayton, P., Pott, A., & Elwakili, N. (2007). Affective forecasting: Why can't people predict their emotions? *Thinking & Reasoning*, *13*(1), 62–80. https://doi.org/10.1080/13546780600872726

Baumert, A., Halmburger, A., & Schmitt, M. (2013). Interventions against norm violations: Dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin*, *39*(8), 1053–1068.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*, 893–897. https://doi.org/10.1037/0022-006X.56.6.893

Bibas, S. (2004). Plea bargaining outside the shadow of trial. *Harvard Law Review*, *117*, 2463.

Bordens, K. S., & Bassett, J. (1985). The plea bargaining process from the defendant's perspective: A field investigation. *Basic and Applied Social Psychology*, *6*(2), 93–110. https://doi.org/10.1207/s15324834basp0602_1

Boster, F. J., Cruz, S., Manata, B., DeAngelis, B. N., & Zhuang, J. (2016). A meta-analytic review of the effect of guilt on compliance. *Social Influence*, *11*(1), 54–67. https://doi.org/10.1080/15534510.2016.1142892

Bowers, J. (2008). Punishing the innocent. *University of Pennsylvania Law Review*, *156*(5), 1117–1179. https://scholarship.law.upenn.edu/penn_law_review/vol156/iss5/1

Caldwell, H. M. (2011). Coercive plea bargaining: The unrecognized scourge of the justice system. *Catholic University Law Review*, *61*(1), 63–96. https://scholarship.law.edu/lawreview/vol61/iss1/2

Callan, M. J., Sutton, R. M., Harvey, A. J., & Dawtry, R. J. (2014). Immanent justice reasoning: Theory, research, and current directions. In M. Zanna & J. Olson (Eds.), *Advances in experimental social psychology* (Vol. 49, pp. 105–161). Academic Press.

Cardenas, S. A. (2023). Charged up and anchored down: A test of two pathways to judgmental and decisional anchoring biases in plea negotiations. *Psychology, Public Policy, and Law*. Advanced online publication. https://doi.org/10.1037/law0000390

Carlsmith, J. M., & Gross, A. E. (1969). Some effects of guilt on compliance. *Journal of Personality and Social Psychology*, *11*(3), 232–239. https://doi.org/10.1037/h0027039

Cialdini, R. B., Darby, B. L., & Vincent, J. E. (1973). Transgression and altruism: A case for hedonism. *Journal of Experimental Social Psychology*, *9*(6), 502–516. https://doi.org/10.1016/0022-1031(73)90031-0

Davis, D., & O'Donohue, W. T. (2004). The road to perdition: Extreme influence tactics in the interrogation room. In W. T. O'Donohue & E. R. Levensky (Eds.), *Handbook of forensic psychology: Resource for mental health and legal professionals* (pp. 897–996). Elsevier Science. https://doi.org/10.1016/B978-012524196-0/50037-1

DeCelles, K. A., Adams, G. S., Howe, H. S., & John, L. K. (2021). Anger damns the innocent. *Psychological Science*, *32*(8), 1214–1226. https://doi.org/10.1177/0956797621994770

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dervan, L. E., & Edkins, V. A. (2013). Innocent defendant's dilemma: An innovative empirical study of plea bargaining's innocence problem. *The Journal Criminal Law & Criminology*, *103*, 1.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*(6), 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self-or social prediction? *Journal of Personality and Social Psychology*, *79*(6), 861–875. https://doi.org/10.1037/0022-3514.79.6.861

Feather, N. T. (1999). Judgments of deservingness: Studies in the psychology of justice and achievement. *Personality and Social Psychology Review*, *3*(2), 86–107. https://doi.org/10.1207/s15327957pspr0302_1

Feather, N. T. (2006). Deservingness and emotions: Applying the structural model of deservingness to the analysis of affective reactions to outcomes. *European Review of Social Psychology*, *17*(1), 38–73. https://doi.org/10.1080/10463280600662321

Feather, N. T., McKee, I. R., & Bekker, N. (2011). Deservingness and emotions: Testing a structural model that relates discrete emotions to the perceived deservingness of positive or negative outcomes. *Motivation and Emotion*, *35*(1), 1–13. https://doi.org/10.1007/s11031-011-9202-4

Freedman, J. L., Wallington, S. A., & Bless, E. (1967). Compliance without pressure: The effect of guilt. *Journal of Personality and Social Psychology*, *7*(2), 117–124. https://doi.org/10.1037/h0025009

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*, 332–346. https://doi.org/10.1037//0022-3514.75.2.332

Gregory, W. L., Mowen, J. C., & Linder, D. E. (1978). Social psychology and plea bargaining: Applications, methodology, and theory. *Journal of Personality and Social Psychology*, *36*(12), 1521–1530. https://doi.org/10.1037/0022-3514.36.12.1521

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. John Wiley.

Guyll, M., Madon, S., Yang, Y., Lannin, D. G., Scherr, K., & Greathouse, S. (2013). Innocence and resisting confession during interrogation: Effects on physiologic activity. *Law and Human Behavior*, *37*, 366–375. https://doi.org/10.1037/lhb0000044

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 451–470. https://doi.org/10.1111/bmsp.12028

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Henderson, K. S., & Levett, L. M. (2018). Investigating predictors of true and false guilty pleas. *Law and Human Behavior*, *42*(5), 427–441. https://doi.org/10.1037/lhb0000297

Inbau, F., Reid, J., Buckley, J., & Jayne, B. (2013). *Criminal interrogation and confessions*. Jones & Bartlett Publishers.

The Innocence Project. (n.d.). *Cases*. Retrieved on July 12, 2023, https://www.innocenceproject.org/all-cases/

Kassin, S. M. (2005). On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist*, *60*(3), 215–218. https://doi.org/10.1037/0003-066X.60.3.215

Kassin, S. M., & Norwick, R. J. (2004). Why suspects waive their Miranda rights: The power of innocence. *Law and Human Behavior*, *28*, 211–221. https://doi.org/10.1023/B:LAHU.0000022323.74584.f5

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276–278.

Kramer, G. M., Wolbransky, M., & Heilbrun, K. (2007). Plea bargaining recommendations by criminal defense attorneys: Evidence strength, potential sentence, and defendant preference. *Behavioral Sciences & the Law*, *25*(4), 573–585. https://doi.org/10.1002/bsl.759

Läkens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., & Caldwell, A. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920951503

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion*. Plenum.

The National Registry of Exonerations. (n.d.). *Search Exonerations*. Retrieved July 12, 2023. https://www.law.umich.edu/special/exoneration/Pages/detaillist.aspx

Marschall, D., Sanftner, J., & Tangney, J. P. (1994). *The State Shame and Guilt Scale*. Fairfax, VA: George Mason University.

Oeberst, A., & Imhoff, R. (2023). Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916221148147

Pardieck, A. M., Edkins, V. A., & Dervan, L. E. (2020). Bargained justice: The rise of false testimony for false pleas. *Fordham International Law Journal*, *44*(2), 469–528. https://ir.lawnet.fordham.edu/ilj/vol44/iss2/4

Perillo, J. T., Crozier, W. E., Pollick, C. L., & Kassin, S. M. (March 2014). *The effect of prior false confession on guilty plea decisions* [Paper presentation]. American Psychology-law Society Conference, *New Orleans, LA, United States*.

Perillo, J. T., & Kassin, S. M. (2011). Inside interrogation: The lie, the bluff, and false confessions. *Law and Human Behavior*, *35*(4), 327–337. https://doi.org/10.1007/s10979-010-9244-2

Redlich, A. D., Bushway, S. D., & Norris, R. J. (2016). Plea decision-making by attorneys and judges. *Journal of Experimental Criminology*, *12*(4), 537–561. https://doi.org/10.1007/s11292-016-9264-0

Redlich, A. D., & Shteynberg, R. V. (2016). To plead or not to plead: A comparison of juvenile and adult true and false plea decisions. *Law and Human Behavior*, *40*(6), 611–625. https://doi.org/10.1037/lhb0000205

Redlich, A. D., Yan, S., Norris, R. J., & Bushway, S. D. (2017). The influence of confessions on guilty pleas and plea discounts. *Psychology, Public Policy, and Law*, *24*(2), 147–157. https://doi.org/10.1037/law0000144

Robinson, M. D., & Clore, G. L. (2001). Simulation, scenarios, and emotional appraisal: Testing the convergence of real and imagined reactions to emotional stimuli. *Personality and Social Psychology Bulletin*, *27*(11), 1520–1532. https://doi.org/10.1177/01461672012711012

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Russano, M. B., Meissner, C. A., Narchet, F. M., & Kassin, S. M. (2005). Investigating true and false confessions within a novel experimental paradigm. *Psychological Science*, *16*(6), 481–486. https://doi.org/10.1111/j.0956-7976.2005.01560.x

Scherr, K. C., & Franks, A. S. (2015). The world is not fair: An examination of innocent and guilty suspects' waiver decisions. *Law and Human Behavior*, *39*, 142–151. https://doi.org/10.1037/lhb0000121

Sheeran, P. (2002). Intention—behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, *12*(1), 1–36. https://doi.org/10.1080/14792772143000003

Schneider, R. A., & Zottoli, T. M. (2019). Disentangling the effects of plea discount and potential trial sentence on decisions to plead guilty. *Legal Criminological Psychology*, *24*(2), 288–304. https://doi.org/10.1111/lcrp.12157

Tangney, J. P., Stuewig, J., & Hafez, L. (2011). Shame, guilt, and remorse: Implications for offender populations. *Journal of Forensic Psychiatry & Psychology*, *22*(5), 706–723. https://doi.org/10.1080/14789949.2011.617541

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tor, A., Gazal-Ayal, O., & Garcia, S. M. (2010). Fairness and the willingness to accept plea bargain offers. *Journal of Empirical Legal Studies*, *7*(1), 97–116. https://doi.org/10.1111/j.1740-1461.2009.01171.x

Turner, B. M., & Schley, D. R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology*, *90*, 1–47. https://doi.org/10.1016/j.cogpsych.2016.07.003

Tyler, T. R. (2000). Social justice: Outcome and procedure. *International Journal of Psychology*, *35*(2), 117–125. https://doi.org/10.1080/002075900399411

U.S. Sentencing Commission. (2022). *2022 Annual report and sourcebook of federal sentencing statistics*. https://www.ussc.gov/sites/default/files/pdf/research-and-publications/annual-reports-and-sourcebooks/2022/2022-Annual-Report-and-Sourcebook.pdf

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wilford, M. M., Sutherland, K. T., Gonzales, J. E., & Rabinovich, M. (2021). Guilt status influences plea outcomes beyond the shadow-of-the-trial in an interactive simulation of legal procedures. *Law and Human Behavior*, *45*(4), 271–286. https://doi.org/10.1037/lhb0000450

Wilford, M. M. & Wells, G. L. (2018). Bluffed by the dealer: Distinguishing false pleas from false confessions. *Psychology, Public Policy, and Law*, *24*(2), 158–170. https://doi.org/10.1037/law0000165

Wilford, M. M., Wells, G. L., & Frazier, A. (2021). Plea-bargaining law: The impact of innocence, trial penalty, and conviction probability on plea outcomes. *American Journal of Criminal Justice*, *46*(3), 554–575. https://doi.org/10.1007/s12103-020-09564-y

Wilson, T. D., & Gilbert, D. T. (2013). The impact bias is alive and well. *Journal of Personality and Social Psychology*, *105*(5), 740–748. https://doi.org/10.1037/a0032662

Yoffe, E. (2017). Innocence is irrelevant. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2017/09/innocence-is-irrelevant/534171/

Zottoli, T. M., Daftary-Kapur, T., Winters, G. M., & Hogan, C. (2016). Plea discounts, time pressures, and false-guilty pleas in youth and adults who pleaded guilty to felonies in New York City. *Psychology, Public Policy, and Law*, *22*(3), 250–259. https://doi.org/10.1037/law0000095