PUNISHMENT IN PUBLIC GOODS GAMES

Torrin M. Liddell

Submitted to the faculty of the University Graduate School in partial fulfillment of the requirements for the degree Doctor of Philosophy in the Cognitive Science program and the Department of Psychological & Brain Sciences, Indiana University
November 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

John K. Kruschke, Ph. D

Edward Hirt, Ph. D

Jerome Busemeyer, Ph. D

Colin Allen, Ph. D

October 31st, 2018

Acknowledgments

I am grateful to all those whom I have worked with in the pursuit of this research, and more generally, in the pursuit of the degree this dissertation caps. This includes the many excellent faculty I have interacted with here at Indiana University, as well as many of my fellow students. In particular, I am grateful for the many fruitful discussions I have had with my fellow lab member, Brad Celestin.

And of course, I am greatly indebted to my Dissertation Committee members, who have been both tremendously helpful in providing advice and feedback, as well incredibly obliging in navigating the dissertation submission and defense process. My advisor, Dr. John K. Kruschke, has taught me more than anyone I have ever met, and the knowledge imparted has shown no signs of abating. It is an understatement to say that I would not be where I am today without his guidance.

Torrin M Liddell

Punishment in public goods games

Punishment is an important method for discouraging uncooperative behavior. This work studies the information used when deciding to apply a punishment, and what punishment to apply. We use a novel design for a public goods game in which a player's actual contribution is a random deviation from their intended contribution, and both the intended and actual contributions are explicitly displayed to all players. This feature lets players detect accidental free riding or accidental high contributing. Multiple types of punishment are studied, including fines, ostracism, and reputation marking. We investigate the effect of a punishment's efficacy for changing behavior on the continued use of the punishment. We investigate the effect of local norms of punishment. We also investigate the effect of the cost of applying a punishment. Our novel design with automated players allows complete experimental control and thus provides the capability to manipulate these factors directly. Bayesian hierarchical models are used for data analysis. Contrary to some pre-existing literature, punishment decisions are found to be flexible, to be responsive to changing conditions, and to emphasize outcomes over intentions only in specific, narrow circumstances. Moreover, we find that the rarely studied punishments of ostracism and reputation marking are quite different from the more often studied fine in how they are utilized, and thus these and other alternative punishments are essential to study in the future.

John K. Kruschke, Ph. D

Edward Hirt, Ph. D

Jerome Busemeyer, Ph. D

Colin Allen, Ph. D

Contents

Introduction

Punishment is an important mechanism for discouraging uncooperative and antisocial behavior. The role that punishment plays in encouraging cooperation has been heavily studied (eg., Fehr & Gächter, 2000; Ostrom, Walker, & Gardner, 1992). However, less work has been done investigating the decision making process of individual punishers when a potentially punishable transgression occurs. This dissertation presents several novel experiments that investigate features of punishment decision making.

The framework: The public goods game

Punishment behavior is often studied in a *public goods game* or *PGG* (e.g., Carpenter, Verhoogen, & Burks, 2005; Fehr & Gächter, 2000; Ostrom et al., 1992; Walker & Halloran, 2004; Yamagishi, 1986). In a public goods game, each round begins with an endowment distributed equally to all players. Players then individually decide how much to contribute to the common pool (i.e., the public good). The common pool value is then multiplied by a constant greater than one (e.g., the pool is doubled) and the total amount of the pool is equally divided among all players. Thus, all players get an equal share of the public good, regardless of how much they contributed to it. This structure implies that the best outcome for an individual is to contribute nothing while others contribute as much as possible to the common pool. Yet the best outcome for the group overall is for all players to cooperate by contributing their entire initial endowment. There is a conflict between any individual's incentive to free ride and the interests of the group as a whole. When an individual places their interests above the group and free riding occurs, other players may want to punish the free rider. The novel studies presented here utilize variations of the PGG to test hypotheses regarding what influences decisions to punish.

Organization of this dissertation

The remainder of this introduction presents select previous works that are especially relevant for contextualizing all of the new experiments presented here, with each chapter also containing brief, chapter specific, literature reviews. The following chapters each present data from one or more original experiments that are relevant to a particular topic in punishment decision making. Chapters 1, 2, 3, and 5 present edited portions of work published in Liddell and Kruschke (2014) on the topics of the punishment of accidents, ostracism as a punishment option, the role of punishment norms, and punishment efficacy, respectively. Chapters 4, 6, and 7 present previously unpublished work on the role of proximate punishment cost, reputation damage as a punishment option, and variations in punishment decision making across the two primary populations sampled (Indiana University undergraduates and Amazon Mechanical Turk participants). Finally, the appendices describe the data analysis methods in detail.

Intention and outcome

When deciding to punish some negative action, we may feel that the intention behind the act ought to be an extremely important factor in the decision to punish. Similarly, the punishment of legitimate and unpreventable accidents feels wrong in some way. However, this account does not necessarily match behavior observed in the lab. Outcome bias occurs when the outcome of an event contributes to the evaluation of an action even when all other aspects of the action (e.g., intention of the actor, reasoning of the actor) are held constant (Baron & Hershey, 1988). When people make a punishment decision, they often exhibit outcome bias. Moreover, they can exhibit an even more extreme behavior pattern we refer to as outcome emphasis, which means that they weigh the actual outcome of the transgression *more strongly* than the transgressor's intended outcome. When there is outcome emphasis, accidental transgressions tend to be punished, but attempted transgressions

that fail to occur tend to be excused. Consider, for example, work performed by Cushman, Dreber, Wang, and Costa (2009) using a "trembling hand" economic game, so named for a scenario involving gunmen with trembling hands who might intend to hit their target but accidentally miss, or who intend merely to scare their target with a close miss but accidentally hit. In this game, one player was given an amount of money to allocate between him/herself and the second player. After allocation, the second player was allowed to respond. This response was either to apply a monetary punishment, give a monetary reward, or do nothing. The "trembling hand" feature of the game was that the allocating player, instead of directly choosing an allocation, chose one of three dice. The three dice had different probabilities of (a) selfishly keeping the entire allocation, (b) fairly splitting it, or (c) generously giving it all away to the second player. One die had a 2/3 chance of being selfish, a second die had a 2/3 chance of being fair, and a third die had a 2/3 chance of being generous (with the other two unspecified allocations having 1/6 probability in all cases). Thus, the allocator could intend to be selfish, fair, or generous, but accidentally roll an unintended outcome. The choice of die was explicitly revealed to the receiving player, so the intention of the allocator was directly observable (at least in terms of what outcome the allocator was attempting to cause). Results showed that when making punishment decisions, participants put much greater weight on the actual outcome (the final allocation) as compared to the intended outcome (the die chosen). In other words, people were willing to punish allocators who were accidentally selfish, despite knowing that the allocator intended to be fair or generous.

Despite the evidence supporting the existence of outcome emphasis in punishment, we have reason to think that outcome emphasis may not extend to all punishments or situations. In the context of our PGG, the intention and outcome distinction corresponds to the intended contribution to the common pool that the player selects and the actual contribution that the player makes. These two values are allowed to differ via the introduction of

a random noise component. Investigation of the relative role of these different sources of information when decisions about punishment are being made pervades all the experiments presented in this dissertation.

Automation

To assess many of the research questions addressed in this dissertation it was efficacious to have complete experimental control of the game environment, and therefore we automated all the players other than the single human participant. Automated players have been used in previous research in the context of PGGs (Barclay, 2006; Suri & Watts, 2011), other economic games (e.g. the prisoner's dilemma in Kiesler, Sproull, & Waters, 1996), and other forms of experimental games (e.g. a blame attribution game in Gerstenberg & Lagnado, 2010). Moreover, there is evidence that computerized players are treated as human in multiple contexts (e.g., Fogg & Nass, 1997; Nass, Fogg, & Moon, 1996; Nass & Moon, 2000).

In all of our experiments, players were instructed that they may be playing against networked or automated players. We did not collect any measures that assessed whether or not participants believed they were playing against automated or human players. For in person experiments, all participants were given the opportunity to present any comments or questions they had regarding the experiment, and participants generally expressed uncertainty regarding whether they were playing against human or automated players. However, the vast majority of participants did not bring up the topic.

Chapter 1: Accidents in the PGG framework

A key innovation for our PGG is that contributions to the public good are affected by a trembling hand. All players see the intended and actual contributions to the public good. We set out to test if outcome emphasis (for clarity, we refer to outcome emphasis in the context of our design as "emphasis on actual contribution") in punishment would occur in a trembling-hand PGG, where the bias would manifest as punishments that emphasize the actual contribution more than the intended contribution. This directly assesses our primary research question of the information utilized when deciding to apply a punishment, especially intention and outcome information.

Experiment 1.1: Assessing emphasis on actual contribution in the automated PGG

Methods. 160 Indiana University (IU) undergraduates participated in the experiment for course credit. Participants were recruited from the human subjects pool of the Department of Psychological and Brain Sciences. We assume the participants were representative of the pool, which is approximately 65% female with ages ranging approximately from 18 to 45 years with a modal age of 19.

Participants were told they would be playing a game while seated at a computer with other players who might be networked people or automated. Each player was referred to by a static single letter label. At the beginning of each round players were given 10 points and allowed to contribute as many points as they wished to a common pool. This contribution was described as an investment in a group venture. Following this choice, noise was applied to the intended contribution to produce the actual contribution. The noise was a random integer chosen uniformly from the set 2, 3, and 4, and then assigned a positive or negative sign with equal chance. This noise pattern (particularly, the lack of 0 noise) was chosen in order for the influence of the intended and actual contribution to be more easily distinguished. Participants were told that this random noise reflected real world contingencies

such as miscommunications or mistakes. This value was added to the intended contribution to produce the actual contribution. Actual contributions could not be below 0 or above 10. Every game had five players, four of which were automated. The actions of the automated players were randomly selected from a pregenerated list of contribution combinations. A complete list is available in the supplementary materials. Each combination consisted of a player who was intentionally low, a player who was accidentally low, a player who was intentionally high, and a player who was accidentally high. The "low" and "high" designations are relative to a baseline of 5. After the participant made her contribution, all of the other players' contributions were displayed in a table on the computer screen. This table also contained the amount paid out to each player from the pool and each player's total gain for the round. Payout was equal to the total amount contributed multiplied by 1.6 and then divided equally among all players. Total gain was equal to the payout from the pool plus any amount the player kept from their initial allocation.

After the contributions and payoffs were displayed, the participant had the opportunity to punish the other players. In phase 1 of the experiment, consisting of 22 rounds, only the participant was given the opportunity to punish other players. Participants were not made aware of the length of this phase, or even that there would be a second phase. Punishment consisted of deducting points from the player, at a cost to the punisher of a quarter point per point deducted from the target. Participants could punish any number of players as long as they did not attempt to spend more points than they gained in the round. An example of the interface participants saw is shown in Figure 1.

In phase 2 of the experiment, again consisting of 22 rounds, the automated players also applied penalties. Participants were told that they were starting a new game with new players, and that the other players could apply penalties. Every automated player applied a penalty to every other player equal to the difference between the other player's actual contribution and the mean actual contribution. The participant applied her penalties without

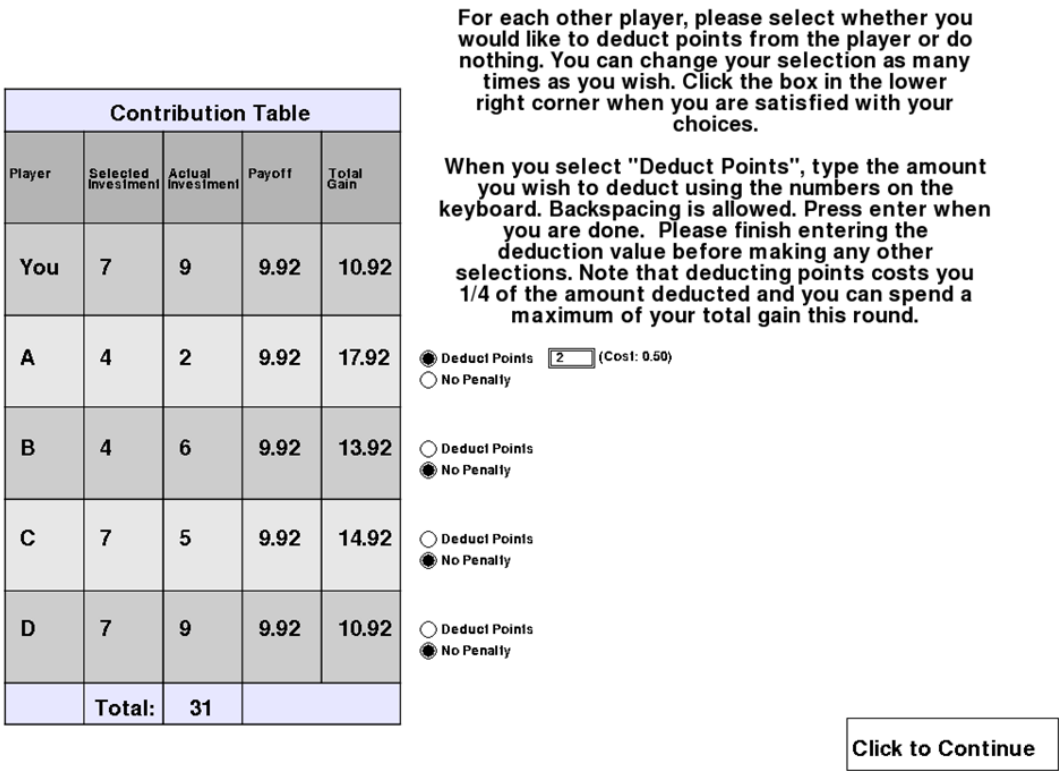


Figure 1. An example of the Experiment 1.1 punishment choice interface.

seeing the other players' penalties. After the participant applied her penalties, a table was displayed that showed the punishments applied by all players to all players, along with the net gain after penalties. The purpose of this two-phase design was to be able to observe the behavior of participants unbiased by the punishment behavior of the automated players (phase 1) and also in the presence of other punishing players (phase 2).

The trembling-hand PGG has several other novelties relative to previous research. In our trembling-hand PGG, punishers are also contributors, unlike in previous work with a different paradigm in which punishers were only responding to the actions of others (Cushman et al., 2009). In the trembling-hand PGG, many rounds are actually played consecutively instead of using the "strategy method" in which hypothetical judgments are solicited from each participant. Finally, in our trembling-hand PGG, the other players are

automated to give us complete control of the game environment.

Results. When deciding to punish, the participant sees three sources of information about herself and the other players, namely the intended contribution, the actual contribution, and the net gain. We are interested in how much each source of information is weighted in the decision to apply a fine. (Note that we are analyzing the probability of applying any fine, not the magnitude of fine applied. We do so to allow comparison with the other punishment types introduced in later chapters that lack magnitude. However, we also performed a linear regression of fine amount with similar results). To model the probability of applying a fine, we used logistic regression on three predictors: the intended contribution of the targeted player, the actual contribution of the targeted player, and the extent to which the targeted player got more net points than the punisher, which we call "indignation." Colloquially, indignation is a sense of anger or annoyance at perceived injustice. This label is a convenient mnemonic for the numerical predictor, but it is not intended to imply that we measured a subjective attitude.

Indignation as we define it here is closely related to the concept of inequity aversion as described by Fehr and Schmidt (1999). In the model utilized by Fehr and Schmidt (1999), inequity is the average difference in payout between a given player and all other players. Inequity so defined is essentially average indignation as defined in our analysis. Thus, our including indignation as a predictor allows the regression model to distinguish the influence of personal inequity (indignation) from the influence of actual contribution.

Bayesian hierarchical logistic regression. The hierarchical model applies logistic regression to the behavior of each individual, and additionally estimated higher-level distributions across the individual regression parameters to describe group-level tendencies. For a full description of the model, see Appendix 1. The important parameters for our purposes are the normalized group-level regression weights, which indicate the relative influence of the three predictors. The regression weights are denoted β_{act} for the actual predictor,

β_{int} for the intention predictor, and β_{indig} for the indignation predictor. The regression weights are "normalized" by dividing each of the raw regression weights by the sum of all the squared regression weights. This normalization across predictors is reasonable because the scales of the three predictors are the same: monetary points. The normalized regression weights represent the values of the raw regression weights relative to one another. This allows easier comparison across regression weights, and in later experiments across conditions. These normalized regression weights are referred to by "beta weights" or just "weights" from here on.

These beta weights represent the relative importance of the given predictor in determining the probability of applying a fine, at the level of the group tendency. A large magnitude beta weight represents that the predictor is relatively important, and a beta weight near zero indicates that the associated predictor is relatively unimportant for predicting the application of a fine. Furthermore, a positive beta weight indicates that a higher value on that predictor produces a higher probability of fining (as would be expected for indignation) whereas a negative beta weight indicates that a higher value of the predictor produces a lower probability of fining (as would be expected for actual contribution and intended contribution).

We estimate the parameters using Bayesian methods (Gelman et al., 2013; Kruschke, 2013, 2015; Kruschke, Aguinis, & Joo, 2012; Kruschke & Liddell, 2018a, 2018b; Ntzoufras, 2009). Bayesian estimation is especially seamless for complex hierarchical models such as the one used here, because it yields a complete posterior distribution of jointly credible parameter values, given the data. There is no need to compute p values from auxiliary sampling assumptions and null hypotheses. We use Markov chain Monte Carlo (MCMC) techniques programmed in R, JAGS (Plummer, 2003) and runjags (Denwood, 2013) to generate representative credible values from the joint posterior distribution. The chains were burned in and checked for convergence, and run long enough to produce an effective sam-

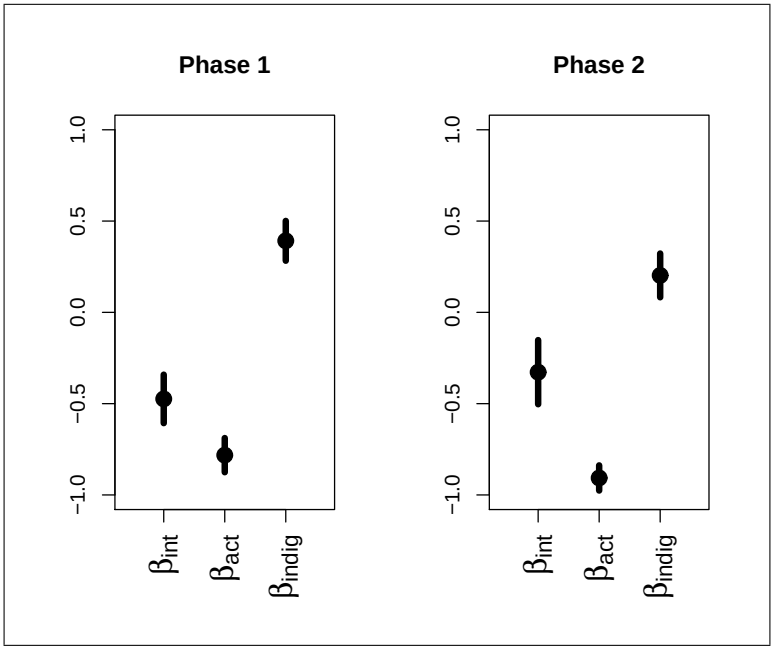


Figure 2. Parameter estimates from Experiment 1.1, showing marginal posterior distributions of the normalized group-level regression coefficients. The vertical black bars indicate the 95% highest density interval (HDI) which contains the most credible 95% of the values, with the point indicating the mean. In both phases, the regression weight on actual contribution is of greater magnitude (more negative) than the regression weight on intended contribution.

ple size (ESS) of at least 10,000 for all of the reported results. This yields a stable and accurate representation of the posterior distribution on the parameters. Except when noted, these features apply to all analyses presented here.

Parameter Estimates. We analyze the data of phase 1 (in which only the participant could apply a fine) separately from the data of phase 2 (in which all players could apply fines). Figure 2 shows the *95% highest density intervals* (HDIs) on the beta weights for each predictor in phases 1 and 2. The 95% HDI contains the 95% most probable parameter values, and is useful as a summary of the posterior distribution, along with the distribution's central tendency. The 95% HDI can also be used as part of a decision rule for rejecting or accepting a null value (Kruschke, 2011, 2013, 2015, 2018). The decision rule uses a *region of practical equivalence* (ROPE) around the null value, which indicates a band of values

that are equivalent to the null for practical purposes. If the HDI falls completely outside the ROPE, the null value is rejected. We will say in this case that the parameter is "credibly" greater than or less than the null value. If the HDI falls completely inside the ROPE, the null value is accepted for practical purposes. In this article, we leave the ROPE tacit, recognizing that the bounds of practical equivalence are not crucial for our claims.

As expected, the regression weights on the intended contribution and actual contribution are negative, meaning that the probability of punishing decreases as intended and actual contributions increase. The weight on indignation is positive, meaning that the probability of punishing increases as indignation increases. This positive weighting suggests that inequity aversion influences punishment in our novel PGG, analogous to previous results in different procedures (Cushman et al., 2009).

We are most interested in the relative weights of intended contribution and actual contribution. It is evident from Figure 2 that the regression weight on actual contribution is of greater magnitude (i.e., more negative) than the regression weight on intended contribution. To quantitatively assess the relative weights of these two predictors, we computed the difference of the regression weights at each step of the MCMC chain. In phase 1, the weight on actual contribution is larger than the weight on intended contribution (mean difference = 0.313, 95% HDI from 0.098 to 0.529), and in phase 2 this difference is even stronger (mean difference = 0.831, 95% HDI from 0.656 to 0.992). Thus, we have shown, for the first time in a trembling-hand PGG, that people deciding to fine weigh actual contributions more heavily than intended contributions.

The relative emphasis on actual contribution increases from phase 1 to phase 2. One possible reason is that participants became more familiar with the task and increased their consistency of responding, allowing the trend to be more clearly expressed. A second possible reason is that behavior in phase 2 reflects mimicking of the automated players, who applied fines based on actual contribution. Experiment 3.1 explores this latter possibility,

and shows that although mimicking may play a role in participants' punishments, heavier weighting of actual contribution is maintained by people even when the automated players punish only on the basis of intended contributions.

Chapter 2: Ostracism as an alternative punishment

Having established in Experiment 1.1 that there is emphasis on actual contribution when deciding to fine in a PGG, we now turn to the alternative punishment of ostracism. Punishments in PGGs and other economic games are usually costly fines (e.g., Cushman et al., 2009; Fehr & Gächter, 2000; Fehr & Schmidt, 1999; Ostrom et al., 1992), as in Experiment 1.1. This type of punishment allows players to deduct resources from another player at a cost to themselves. However, another real-world punishment is ostracism. Ostracism entails a refusal of repeat business with the punished party, and by definition has no immediate cost. Ostracism prevents any future transgressions from the punished party. This type of punishment can also motivate cooperation in public goods games (Cinyabuguma, Page, & Putterman, 2005; Maier-Rigaud, Martinsson, & Staffiero, 2010; Masclet, 2003), but it has been studied relatively rarely.

Baumard has suggested that ostracism is much more representative of everyday punishment than costly fine (Baumard, 2010, 2011; Baumard, André, & Sperber, 2013). He cited anthropological literature to argue that in the hunter-gather societies representative of the environments under which humans evolved, costly punishment is exceedingly rare. Furthermore, he argued that human cooperation can be explained by partner choice alone, which is simultaneously inexpensive compared to a costly punishment and prevents any future transgressions This account is consistent with research on non-human animals that indicates that costly punishment is quite rare and that ostracism is much more frequently observed (Raihani & McAuliffe, 2012; Stevens, Cushman, & Hauser, 2005).

There is also evidence from game-theoretic computer simulations that exclusion may be more conducive to the evolution of cooperation than other forms of punishment. The simulations of Sasaki and Uchida (2013) assumed that ostracism of a freerider resulted in immediate benefits for the remaining cooperative group members because of less dilution

of group output in subsequent rounds. On the other hand, costly fines produce no direct benefit if the punished individual does not increase his contribution in subsequent rounds.

There is also reason to think that emphasis on actual contribution may not extend to ostracism. Punishers may be less willing to lose a person who, based on their intentions, is likely to be cooperative in the future, even if their intentions did not yield cooperative behavior in the present encounter.

Experiment 2.1: Assessing emphasis on actual contribution in ostracism

Given that ostracism has important structural differences from costly fine and that ostracism may be an especially important form of punishment in the real world, we wished to directly compare both forms of punishment in our PGG paradigm. We hypothesized that decisions to ostracize would place more emphasis on intended contribution than decisions to fine, because we expected that participants would be less willing to lose a well-intentioned partner in future rounds because of an accidental outcome. To test this hypothesis we conducted an experiment very similar to Experiment 1.1 that included ostracism as a punishment option.

Methods. 351 IU undergraduates participated in the experiment for course credit. Participants were recruited via the IU Psychological and Brain Sciences human subject pool, with demographics as reported for Experiment 1.1.

Like the procedure in Experiment 1.1, participants played a two-phase public goods game with four automated opponents. The automated behavior was randomly selected from the same pre-existing distribution as was used for Experiment 1.1. The contribution procedure was identical to Experiment 1.1. However, the punishment process had an important elaboration in that punishment could consist of imposing a costly fine as in Experiment 1.1 *or* ostracizing the player from the game at no cost to the punisher. An example of the Experiment 2.1 punishment interface is shown in Figure 3. Only one of these pun-

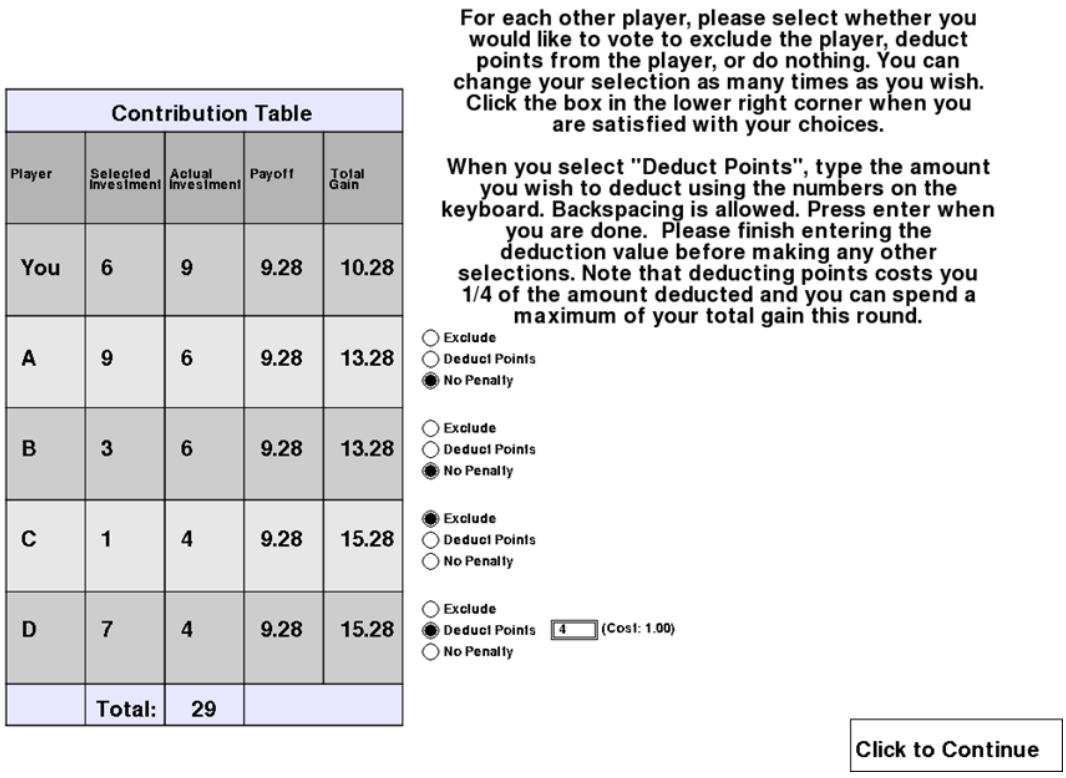


Figure 3. An example of the Experiment 2.1 punishment choice interface.

ishments could be imposed on any one player. If an automated player was excluded, the player would be replaced by a new automated player on the next round. If the participant was excluded by the automated players, the participant would experience a 15-second time out while a message was displayed that described that the system was searching for a new game. The participant would then be put into a new round with all new automated players. The exclusion did not change the total number of rounds played.

In the context of the trembling hand PGG, we refer to ostracism as "exclusion." We expected this term to convey more clearly the nature of this punishment to participants, as ostracism might connote reputation effects that were not explicit in the game. Furthermore, we expected that the term "exclusion" would be more familiar and easy to understand to the average participant than "ostracism." Thus when referring to the ostracism punishment

that can be enacted in the trembling-hand PGG we refer to "exclusion," and when referring to theoretical results about punishment we refer to "ostracism."

As before, in phase 1 of the experiment (the first 22 rounds) only the participant could apply penalties, whereas in phase 2 of the experiment (the last 22 rounds) the automated players could also apply punishments. The automated players excluded a player if her intended contribution was at least 2 points lower than the mean intended contribution. If a player did not meet the exclusion criterion, the automated players applied costly fines using the same punishment rule as in Experiment 1.1.

Results. To analyze the punishment behavior in Experiment 2.1 we again use a Bayesian hierarchical model that predicts the probability of each punishment choice given the value of the three predictors: actual contribution, intended contribution, and indignation. However, now the analysis concerns a trinary choice, not a binary one. To handle the trinary choices, we use a conditional logistic regression that predicts two choice probabilities. The first is the probability of applying exclusion versus not applying exclusion. The second is the probability of applying a fine, given that no exclusion was applied. The analysis is a *conditional* logistic regression because this second probability is conditional on the first choice (exclusion) not occurring.

A traditional analysis for n-ary choice data is multinomial logistic regression, which models the probabilities of all choices without conditionalizing on any one of them. We instead use conditional logistic regression because the multinomial model assumes the independence of irrelevant alternatives (Luce, 1959, 2008), which we do not have reason to believe applies to our data. A full description of the model as well as more detailed discussion of the modeling choices are available in the appendices.

Bayesian hierarchical conditional logistic regression. A detailed description of the model is available in the appendices. Again, the primary parameters of interest are the normalized group-level beta weights just as in Experiment 1.1, but each of the three predic-

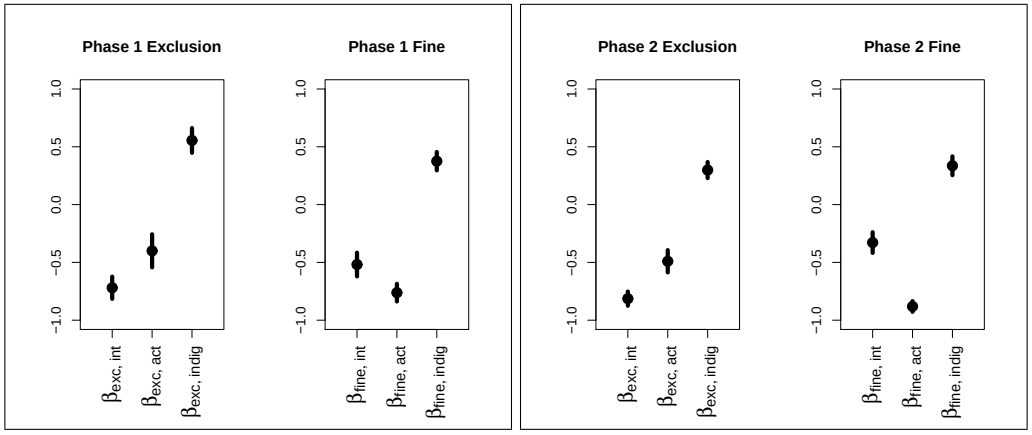


Figure 4. Results from Experiment 2.1, showing 95% HDIs for the posterior distributions of beta weights for exclusion (left side of each panel) and fining (right side of each panel). Notice that for exclusion, the magnitude of the beta weight on intended contribution is larger (i.e., more negative) than on actual contribution, but for fining the opposite is true.

tors now has two sets of beta weights. For the probability of exclusion, the three weights are denoted $\beta_{exc,act}$, $\beta_{exc,int}$, and $\beta_{exc,indig}$. The second group of beta weights predicts the probability of applying a fine given no exclusion occurred, and are analogously denoted $\beta_{fine,act}$, $\beta_{fine,int}$, and $\beta_{fine,indig}$. These beta weights are interpreted just as before, but now each beta weight concerns both a specific predictor and a specific punishment. We again use MCMC techniques to generate 20,000 representative credible values from the joint posterior distribution on the 2,825 parameters in each phase (see Appendix 3 for an analysis of specific rounds). The effective sample size for all results reported below was at least 10,000.

Parameter Estimates. Figure 4 shows the 95% HDIs of the beta weights. As expected intuitively and as found in Experiment 1.1, the weights on the intended contribution and actual contribution are negative, and the weights on indignation are positive.

The two sides of each panel of Figure 4 show the weights for excluding and fining. Importantly, notice that for excluding, the weight on the intended contribution is of greater magnitude (more negative) than the weight on the actual contribution. However, for fining,

the opposite is true, as it was in Experiment 1.1. Thus, we have shown for the first time that people emphasize intended contributions more than actual contributions when deciding to ostracize, but emphasize actual contributions more than intended contributions when deciding to fine. Quantitative analysis verifies the apparent differences in Figure 4, as detailed in the following paragraphs.

In phase 1, consider the weight on intended contribution, comparing across exclusion and fine (i.e., $\beta_{fine,int}$ versus $\beta_{exc,int}$): the mean difference is 0.202, with 95% HDI from 0.058 to 0.345. Consider the weight on actual contribution, comparing across exclusion and fine (i.e., $\beta_{fine,act}$ versus $\beta_{exc,act}$): the mean difference *in the opposite direction* is 0.365, with 95% HDI from 0.198 to 0.524. Focus now on the weights for exclusion (i.e., $\beta_{exc,int}$ versus $\beta_{exc,act}$): the magnitude of the weight on actual contribution is *less extreme* than the weight on intended contribution, with a mean difference of -0.320, 95% HDI from -0.539 to -0.113. Focusing on fines (i.e., $\beta_{fine,int}$ versus $\beta_{fine,act}$), the weight on actual contribution is *more extreme* than the weight on intended contribution, with a mean difference of 0.247, 95% HDI from 0.072 to 0.416. The same differences are even more pronounced in phase 2.

In phase 2, consider the weight on intended contribution, comparing across exclusion and fine: the mean difference is 0.491, with 95% HDI from 0.382 to 0.599. Consider the weight on actual contribution, comparing across exclusion and fine: the mean difference, again in the opposite direction, is 0.392, with 95% HDI from 0.284 to 0.498. Focus now on the weights for exclusion: the magnitude of the weight on actual contribution is again less than the weight on intended contribution, with a mean difference of -0.326, 95% HDI from -0.480 to -0.171. Focusing on fines, the weight on actual contribution is again more than the weight on intended contribution, with a mean difference of 0.557, 95% HDI from 0.428 to 0.682.

These results verify again that actual contributions are weighed heavily when consid-

ering to punish by fining, but the results show that intended contributions are weighed heavily when considering to ostracize. However, as in Experiment 1, the trends appear to be stronger in phase 2 than in phase 1. As previously discussed, there is the possibility of participants mimicking the punishment behavior of the automated players in phase 2, as the automated players did focus on intention information for exclusion and actual contribution information for fining. However, it is important to note that even in phase 1, when no automated-player penalties were occurring, the pattern that favored actual contribution for fines and intended contribution for exclusion was present. We directly investigate the possible effect of mimicking in Experiment 3.1.

Chapter 3: Norms of punishment

Social norms and punishment are strongly intertwined. Norms set the bar for what is worthy of punishment and what is not (Carpenter et al., 2005; Fehr & Fischbacher, 2004). We are concerned with norms that establish which trangressions are punishable, and norms for what type of punishment to apply. Some previous research explored preexisting norms spontaneously used by individuals in experimental situations. For example, Carpenter and Matthews (2009) were able to estimate the punishment norm participants used in an economic game regarding decisions to punish or not, finding that within one's own group players compare contributions to a high absolute threshold (insensitive to group average), and players that fail to meet this threshold are punished. There are many examples of variations in punishment behavior in laboratory games across cultures (Henrich et al., 2005) including the especially peculiar case of antisocial punishment (Herrmann, Thöni, & Gächter, 2008). Furthermore, there are many examples of variation in norms of punishment in the real world. Studies have found regional and cultural differences in endorsement of corporal punishment of children (Flynn, 1994; Lansford & Dodge, 2008). Attitudes towards the death penalty have fluctuated greatly in the United States, ranging from 42% supporting capital punishment in 1966 to an all time high of 80% in 1996 (Jacobs & Carmichael, 2002; Jones, 2013; Zeisel & Gallup, 1989). More recently, punishments intended to humiliate or shame the offender, such as spending time publicly wearing a sign detailing one's crime, have been controversially reintroduced in some American courts. Public humiliation is a form of punishment that some legal scholars have argued is acceptable under our punishment norms, whereas others argued the opposite (Book, 1999; Kahan, 1996, 2006; Whitman, 1998).

Clearly, social norms play an important role in punishment. We report a new experiment that investigates the interplay of punishment norms and punishment type. We are interested

in the degree to which punishment is influenced by the social norm, and if this influence depends upon the type of punishment under consideration.

Experiment 3.1: Testing acquisition of punishment norms

In the experiment described in this chapter, we investigated multiple hypotheses. For this chapter, we focus on one aspect: the potential acquisition of punishment norms within the scope of the PGG. Recall that in the second phases of the Experiments 1.1 and 2.1, trends became more pronounced. This change could have been due to familiarity with the paradigm and stabilization of response tendencies, or it could have been caused by participants mimicking the punishment tendencies of the automated players. To test this second possibility, we introduced two new punishment rules for the automated players, and we randomly assigned subjects to experience one of the two rules. One rule based punishments only on actual contribution, ignoring intended contribution. Under this punishment rule, the automated players would exclude another player if her actual contribution was 2 or less, otherwise the player would be fined in an amount of how much her actual contribution was less than 8 (with a small amount of random noise applied). If the player contributed at least 8 points, no punishment was applied. The other rule based punishments only on the intended contribution, ignoring the actual contribution, using the same numerical criteria. If participants mimic the behavior of other players, then participants in the two groups should differently weigh actual and intended contributions.

Methods. 258 IU undergraduates participated in the experiment for course credit. Participants were recruited via the IU Psychological and Brain Sciences human subject pool, which has demographics as described previously. Participants were randomly assigned to one of the automated-player punishment rules (actual-contribution focused or intended-contribution focused), resulting in approximately 129 subjects per combination.

Because this experiment is directly interested in the influence of punishments norms,

it did not include an initial phase during which only the participant punished. Instead, all players, including automated ones, were given the full range of punishment options throughout the entirety of the game, which lasted 30 rounds. In addition, Experiment 3.1 also involved several minor changes to increase the feeling of playing with real people. All players were labeled on screen with a random name (instead of a single letter). The names were drawn from the 500 most popular baby names, for males and females, at the United States Social Security baby name data base (http://www.ssa.gov/oact/babynames/). The automated-player contribution and punishment choices had realistic timers before they were displayed on screen, such that they appeared in an asynchronous cascade after the participant entered her intended contribution or punishment.

Results. We use Bayesian conditional logistic regression as in Experiment 2.1. However, we analyze the behavior in each of the two punishment norm conditions separately.

To assess the effect of the automated-player punishment norms, we performed a separate conditional logistic regression analysis on each of the two punishment-norm groups. Because it takes exposure to several examples to experience and learn the punishment norms of the other players, we exclude the initial 10 rounds from the analysis, using the remaining 20 rounds.

Figure 5 plots the beta weights for the groups who experienced automated players punishing on the basis of actual or intended contribution. Notice that the beta weights for fining are similar across the conditions, but the beta weights for excluding are different across the two conditions. The exclusion decisions in the actual-focused condition shows more emphasis on actual outcome than in the intention-focused condition.

To quantitatively assess differences in the beta weights on intended contribution and actual contribution, we subtracted each weight in the intention-focused condition from its corresponding weight in the actual-focused condition. For fining, there was no major difference in the weight on intended contribution (mean difference = -0.085, 95% HDI

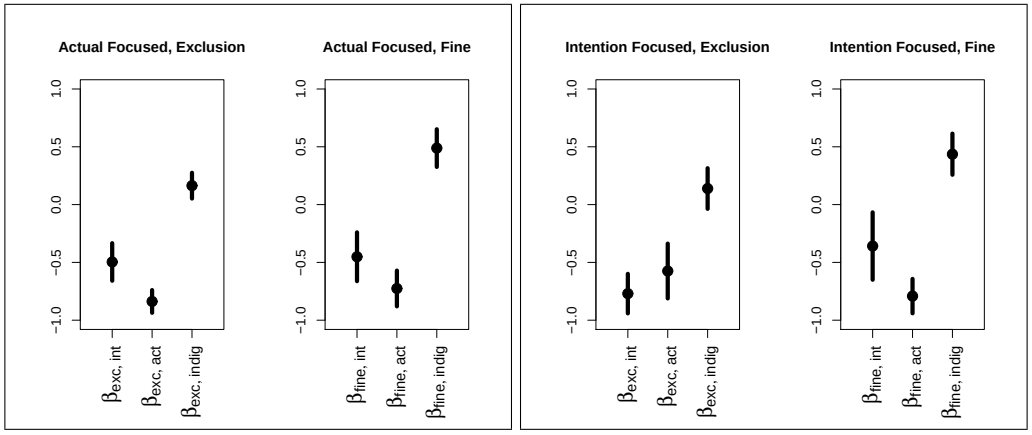


Figure 5. Results of Experiment 3.1 for automated players punishing on the basis of actual contribution (left panel) or on the basis of intended contribution (right panel). Notice that the beta weights for fining (right side of each panel) are similar across the conditions, but the beta weights for excluding (left side of each panel) are different across the two conditions. The exclusion decisions in the actual-focused condition shows more emphasis on actual contribution than in the intention-focused condition.

from -0.457 to 0.271) or on actual contribution (mean difference = 0.071, 95% HDI from -0.156 to 0.287). In contrast, for excluding there was a difference in weights across the two conditions for both intended contribution (mean difference = 0.282, 95% HDI from 0.028 to 0.518) and actual contribution (mean difference = -0.263, 95% HDI from -0.528 to -0.011).

When these results are compared to the first phase of Experiment 2.1 (see Figure 4), where there was no automated player to mimic, a clear pattern emerges. First, fining consistently emphasizes actual contributions, regardless of the punishment norms of the other players. Second, excluding seems to emphasize intended contribution by default, but can be changed to mimic the punishment norms of the group.

Chapter 4: The role of cost

In Experiment 2.1, we compared ostracism and costly fine and concluded that in this paradigm the two punishment types had distinctly different patterns of punishment decision making. However, these two punishment types differed on more than just the punishment type; fines were costly but exclusion was free of any direct cost. It is possible that this cost difference contributed to the differences observed between the two types of punishment. Previous research has investigated the role of cost in the efficiency or efficacy of punishment in promoting cooperation (e.g., Balliet, Mulder, & Van Lange, 2011; Nikiforakis & Normann, 2008; Ohtsuki, Iwasa, & Nowak, 2009), but to our knowledge no work has investigated how cost affects the behavior of punishers with regards to the information utilized in punishment decisions. This section details two separate experiments that investigate directly the role of cost in punishment decision making, in the same PGG framework.

Experiment 4.1: Comparing cost-free punishments

This experiment replicated Experiment 2.1 with one important change; fines were no longer costly, and were a fixed amount of 4 points instead of the amount being selectable from a range. Thus, other than the consequences enacted on the punished individual, the two punishment choices had identical properties.

Methods. 50 IU undergraduates participated in the experiment for course credit. Participants were recruited via the IU Psychological and Brain Sciences human subject pool, with demographics as reported for Experiment 1.1.

As in Experiment 2.1, participants played a two-phase public goods game with four automated opponents with automated behavior being pulled from the same pre-existing distribution as was used for Experiments 1.1 and 2.1. Punishment could again consist of imposing fine or ostracizing the player from the game, both at no cost to the punisher. Imposing a fine always consisted of taking 4 points from the punished player. Figure 6 shows

This table shows your contribution and payoff, along with the contributions and payoffs of all the other players. From this information, please select what you think are appropriate punishments, if any.

Player	Intended Contribution	Actual Contribution	Payout	Gain	
You	7	10	9.6	9.6	
A	7	4	9.6	15.6	○ No Punish ○ Fine: -4pts ○ Exclude
B	9	6	9.6	13.6	○ No Punish ○ Fine: -4pts ○ Exclude
C	3	6	9.6	13.6	○ No Punish ○ Fine: -4pts ○ Exclude
D	1	4	9.6	15.6	○ No Punish ○ Fine: -4pts ○ Exclude

Submit Answers

Figure 6. An example of the Experiment 4.1 punishment choice interface.

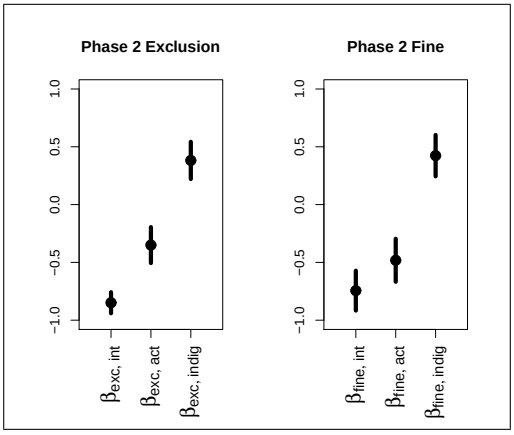


Figure 7. Results of Experiment 4.1. We only report the Phase 2 results here, as the Phase 1 results are qualitatively similar and the small sample size for this study makes fine-grained comparison across phases impossible.

an example of this new interface. All other aspects of the punishment and contribution procedures were identical to Experiment 2.1, except that if the participant was excluded by the automated players, the participant would experience a 9-second time out instead of a 15-second time out.

As before, in phase 1 of the experiment (the first 22 rounds) only the participant could apply penalties, whereas in phase 2 of the experiment (the last 22 rounds) the automated players could also apply punishments. The automated players excluded a player if her intended contribution was at least 2 points lower than the mean intended contribution. If a player did not meet the exclusion criterion, the automated players applied the 4 point fine if her actual contribution was at least 2 points lower than the mean actual contribution.

Results. We use Bayesian conditional logistic regression to analyze punishment behavior, identically to Experiment 2.1.

Figure 7 shows the 95% HDIs of the beta weights. As expected intuitively and as found in Experiment 1.1, the weights on the intended contribution and actual contribution are negative, and the weights on indignation are positive. And as in Experiment 2.1, intention is weighted higher for exclusion than for fines, and actual contribution is weighted higher

for fines than for exclusion, at least in terms of the modal estimates. The posterior HDIs for these comparisons are not reported, as due to the small sample size these differences lack the precision necessary to make substantive conclusions. Despite this, these results were notable due to the relative weight on intention and actual contribution for fines punishment: unlike all previous results, intention is weighted higher than actual contribution for decisions to fine. Thus, although the across-punishment comparison is replicated (at least in direction), the within-punishment assessment of emphasis on actual contribution for fining is not replicated. This motivated us to more directly test the effect of punishment cost in Experiment 4.2.

Experiment 4.2: Direct comparison of costly and cost-free punishments

Experiment 4.2 replicates the general design of Experiment 4.1 but varies both the punishment available to participants, as well as whether the punishment had a direct cost to the punisher. Thus, there were two factors (punishment type and punishment cost) and two conditions per factor (fine/exclusion and costly/cost-free) yielding four conditions. All conditions were entirely between-subject, that is, the punishment option and cost stayed constant throughout the experiment.

Methods. We recruited 130 participants via Amazon Mechanical Turk. Participants were paid $1.20 for their participation. The Amazon Mechanical Turk population is diverse with relatively good data quality in comparison to other convenience sampling methods (Paolacci & Chandler, 2014). Participants were randomly assigned to one of four conditions: costly fine, costly exclusion, free fine, or free exclusion. Each participant only had a single punishment available to them, determined by condition. This punishment could either be costly (1 point per punishment applied) or cost free.

The procedure of Experiment 4.2 was identical to Experiment 4.1 except as noted here. Fining and excluding were identical to their counterparts in Experiment 4.1. Figure 8 shows

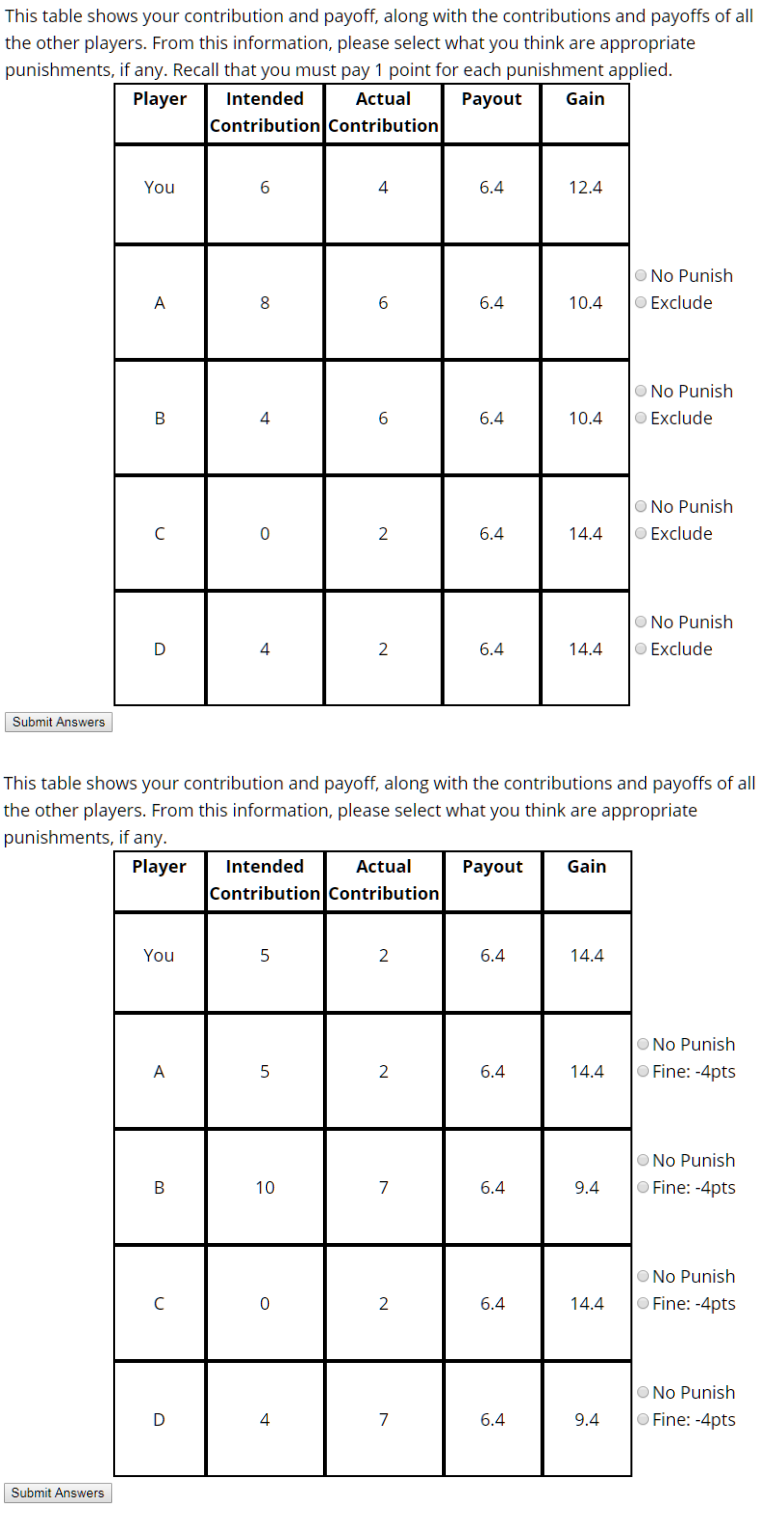


Figure 8. Two example screen shots from of the Experiment 4.2 punishment choice interface, with the top screen showing an example from the costly exclusion condition, and the bottom screen showing an example from the free fine condition.

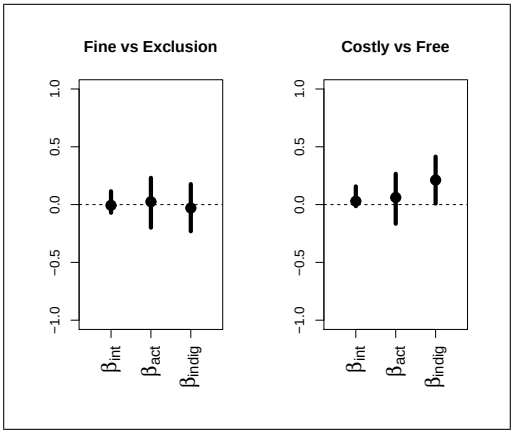


Figure 9. Results of Experiment 4.2. We directly compare across the two factors. The left plot compares average weights in the two fine conditions versus average weights in the two exclusion conditions. The right plot compares average weights in the two costly conditions versus average weights in the two cost-free conditions. Note that unlike previous plots of this type, these are HDIs on the differences between beta-weights. Thus values near zero, as shown in most of the comparisons here, indicate small or no differences associated with that condition axis for that predictor weight.

two example punishment interfaces, showing examples of both punishment type conditions and both cost conditions.

As before, in phase 1 of the experiment (shortened to the first 15 rounds to accommodate the online format) only the participant could apply penalties, whereas in phase 2 of the experiment (also shortened to 15 rounds) the automated players could also apply punishments. The automated players had the same punishment option as the participant, determined by condition. Regardless of condition, the automated players applied a punishment when the target player had an average of intended and actual contribution that was less than 5.

Results. As each participant only had a single punishment available, we used a Bayesian logistic regression to analyze punishment behavior, as described in Experiment 1.1.

Figure 9 shows the 95% HDIs of comparisons across the two factors. Contrary to our hypotheses, and the previous results of Experiments 1 through 3, there were little differ-

ences across either punishment type or punishment cost. All conditions showed the same general pattern of high intention weight, and small weights of actual contribution and indignation. The one potential exception regards indignation; in the costly conditions indignation played a slightly greater role in predicting punishment applications, and this difference was marginally non-zero (95% HDI from 0.01 to 0.42). Thus we have some evidence that cost may have a role in punishment decision making, in that it may draw attention to potential inequity. However, we also have strong evidence of behavior greatly different from previous work that is not directly attributable to experimental changes. Chapter 7 discusses the possibility of both individual and sample differences in emphasis on actual contribution in punishment, with direct replications of Experiments 1.1 and 2.1 in Amazon Mechanical Turk samples.

Chapter 5: The role of reputation

So far we have investigated two punishment types: fine and exclusion. However, reputation damage is another type of punishment that may be of particular relevance due to its important role in cooperation via indirect reciprocity (Mohtashemi & Mui, 2003; Nowak, 2006; Sommerfeld, Krambeck, & Milinski, 2008; Wang, Wang, Yin, & Xia, 2012) as well as the ubiquity and importance of reputation management in the marketplace (Gertsen, van Riel, & Berens, 2006; Rhee & Valdez, 2009) and daily lives of individuals (Kurland & Pelled, 2000; Madden & Smith, 2010). By reputation influence, we mean anything from informal gossip to references to the formalized public ratings common in online commerce. Due to the frequency of reputation influence as a potential response to undesired behavior, it is essential that we investigate it if we are to understand punishment behavior generally. Moreover, this type of response has been little studied in the punishment literature. We implemented this option in the context of our PGG framework in order to assess how reputation ratings are made, as well as how the introduction of a reputation system affects the other punishment behaviors available.

Experiment 5.1: Reputation damage and enhancement

Experiment 5.1 is very similar in concept to Experiment 2.1 with the addition of a third option when participants are given the ability to respond. However, we also changed many structural features of the game in order to increase understanding and engagement, as well as to accommodate the addition of a reputation system.

Methods. 505 IU undergraduates participated in the experiment for course credit. Participants were recruited via the IU Psychological and Brain Sciences human subject pool, with demographics as reported for Experiment 1.1. 43 participants were excluded due to failing to reach the performance threshold in the maximum round length of 40 rounds, for a total of 462 participants. We discuss this performance threshold in more detail below.

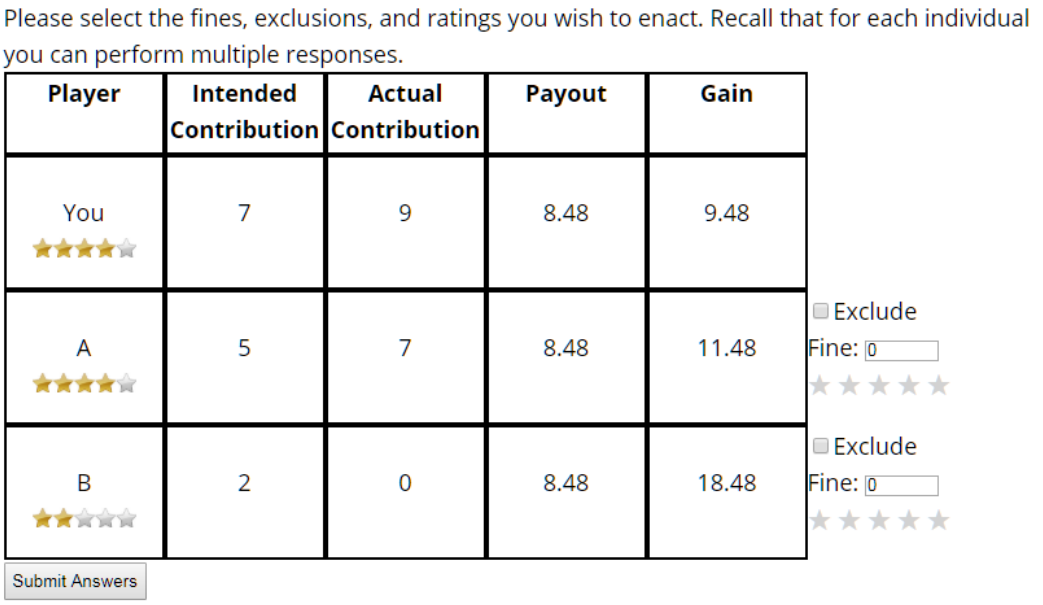


Figure 10. An example of the Experiment 5.1 punishment choice interface.

The primary change is the addition of a possible response providing a rating from 1 to 5 "stars" during the punishment portion of the experiment. The rating history determines a rating displayed below each player's designation consisting of an average of all previous ratings, in a manner similar to how products and sellers are rated on many online storefronts. At the start of the game, or when a new player is added to the game, no rating is displayed for that player. When ratings are received, the average of all ratings received are displayed, rounded to the nearest half-star.

Punishments were not mutually exclusive. That is, more than one punishment could be applied. Moreover, all punishments were cost free, meaning that performing multiple responses did not cause payouts to decrease.

We also made several changes in order to increase engagement with and understanding of the experiment. We reduced the number of automated players to 2, for a total of 3 players, in order to decrease the amount of information simultaneously presented that

participants had to understand and react to. In order to increase motivational salience of succeeding at the game, we changed from a fixed trial length to a target point total. Specifically, in order to complete the experiment, participants needed to reach a total gain of 250 points. This means that better performance leads to fewer rounds played and faster completion of the experiment. As this experiment was completed entirely in the Indiana University Psychological and Brain Sciences subject pool, better performance meant less time spent on the task for the same amount of course credit. If this performance threshold of 250 points was not reached by the end of round 40, the experiment was concluded and that participant's data were excluded from the analysis. As described above, fewer than 10% of participants failed to reach this threshold.

The structural changes described so far necessitated some radical changes in the behavior of the automated players. Automated contributions started at a middling contribution value of 4 points when a player joined the game, with up to a 1 point noise in either direction possible. This tendency was decreased by 2 if no responses were made to that player. If a fine was applied to the automated player, their contribution tendency was increased by the amount of the fine divided by 3, rounded to the nearest integer. Finally, if a high (greater than 3) rating was received, the player maintained their contribution level, unless they were also fined in which case they increased their tendency as described previously.

The automated players also applied punishments and ratings throughout all rounds. However, so as to encourage participant punishment and ratings, they only did so to the participant. Automated players compared the average of the participant's intended and actual contributions to set thresholds. If the average of these two values was less than 1.5, the participant was excluded (with consequences as described previously). If the average of these two values was less than 6, automated players applied a fine equal to the amount the average was exceeded by 6, plus a uniform random integer between -2 and 2. This random adjustment could reduce the fine to 0. Finally, the automated players always provided a

star rating from 1 to 5 based on which of five ranges the average value fell in, from lowest to highest: [0:3), [3,5), [5,7), [7,9), and [9,10].

Results. The analysis presented here differs from the analyses presented in other chapters in several important ways.

The first substantive difference is the addition of a fourth predictor, and thus a fourth set of beta-weights: the existing reputation of the targeted player. This pre-existing reputation is on a 1 to 5 scale rounded to the nearest 0.5 to match the displayed star rating. Players that have no pre-existing reputation were excluded from all analyses presented here. This occurred at the start of the game, as well as any round where a player had yet to have received a rating from the other players. All predictors were standardized to a mean of zero and standard deviation of 1 so that the weights were comparable across the different scales of "points" and the 1 to 5 reputation scale. These standardized weights are also normalized, consistent with previous analyses, but this normalization does not affect the substantive conclusions presented.

Another change is needed in order to account for the non-exclusivity of responses. To do so, we modeled each response separately. Exclusion and fining behavior are both analyzed using separate logistic regression analyses, just as fining behavior was when it was the only option available in Experiment 1.1, with the addition of the fourth reputation predictor as described above.

The unique features of rating behavior means it requires a distinct approach we have not used previously. Rating behavior has two separate components that differ in structure: the decision to provide a rating, and the decision as to what rating ought to be provided. Unlike in the case of fining where the secondary decision (amount of fine) is one of magnitude, the choice of rating changes the valence of the behavior. That is, any fine amount is always a negative valence punishment, whether it is 1 point or 10 points, but a 1 star rating is categorically different from a 5 star rating. Thus it is essential to model both the choice to

apply a rating as well as the choice of rating given.

The choice to provide a rating is analyzed using a logistic regression just as exclusion and fine, but with one elaboration. We had reason to suspect that the probability of applying a rating may not be monotonic with each of the predictors. For instance, a low intention may be associated with a high probability of (a likely low) rating, and a high intention may be associated with a high probability of (a likely high) rating, with middling intention having low probability of rating. This potential "U-shape" is not accommodated by linear predictors. To allow for this curvature, we add an additional quadratic component for each predictor.

As the ratings are ordinal, we model the choice of rating via an ordered-probit regression with parameters again estimated via Bayesian methods (for benefits of this approach, see Liddell & Kruschke, 2018). This approach models the 1 to 5 ratings as ordered responses that are not necessarily equidistant from each other, as a function of an underlying continuous distribution with a mean defined by the predictors. A detailed description of this model is presented in Appendix 4.

Figure 11 summarizes the results of the experiment for exclusion and fine. Exclusion demonstrates heavy emphasis on intended contribution, with a smaller negative weight on actual contribution, and a smaller positive weight on indignation, all as has been typically observed in previous experiments. There is a small negative weight of pre-existing reputation rating, indicating that higher previous ratings are associated with lower likelihood of exclusion.

Conversely, the weights for fining are quite distinct from what has been seen previously. There are moderate negative weights on actual contribution and intended contribution, with no credible difference between the two unlike previous experiments. The effect of indignation is very large in comparison to all other weights, also unlike previous experiments. Finally the weight on reputation is slightly negative, indicating again that high reputation

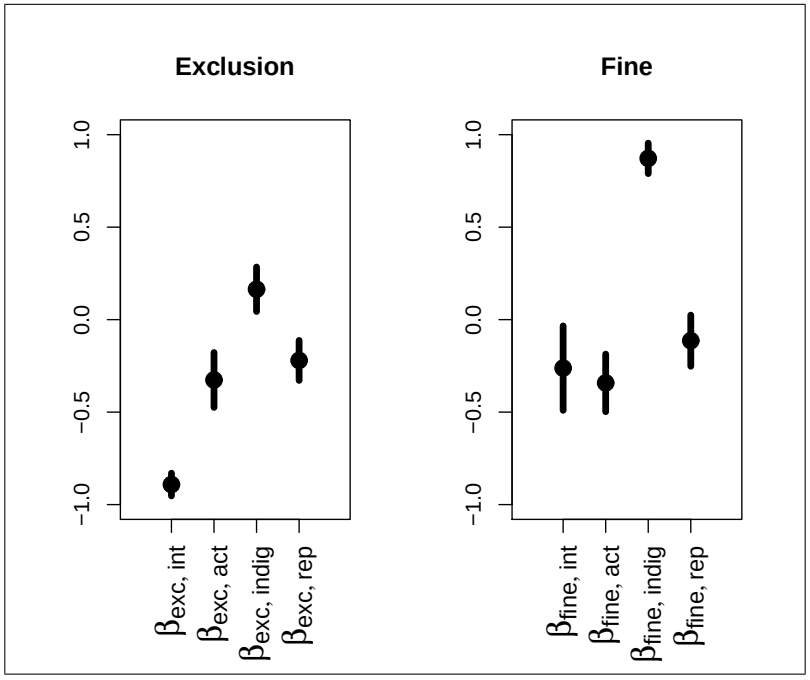


Figure 11. Results from Experiment 5.1 summarizing the fine and exclusion behavior, showing 95% HDIs for the posterior distributions of beta weights.

is associated with lower punishment probabilities. Taken together, it appears that fines are primarily used in scenarios where the target player has done much better than the participant, with the other available information playing much smaller roles. It may be that in the presence of several non-exclusive options, fines become relegated to the role they are most uniquely suited to: correcting personal inequity.

Figure 12 contains the two components of rating behavior. The left panel contains the linear component of the weights for probability of applying a rating. Note that the interpretation of these linear coefficients only applies to each predictor at their mean, and a more complete interpretation including the quadratic trends is shown in Figure 13. The primary predictor of applying a rating is low intended contribution, suggesting that ratings are primarily being used as a punishment in response to low intended contributions. Previous reputation and actual contribution have smaller negative weights, further indicating the

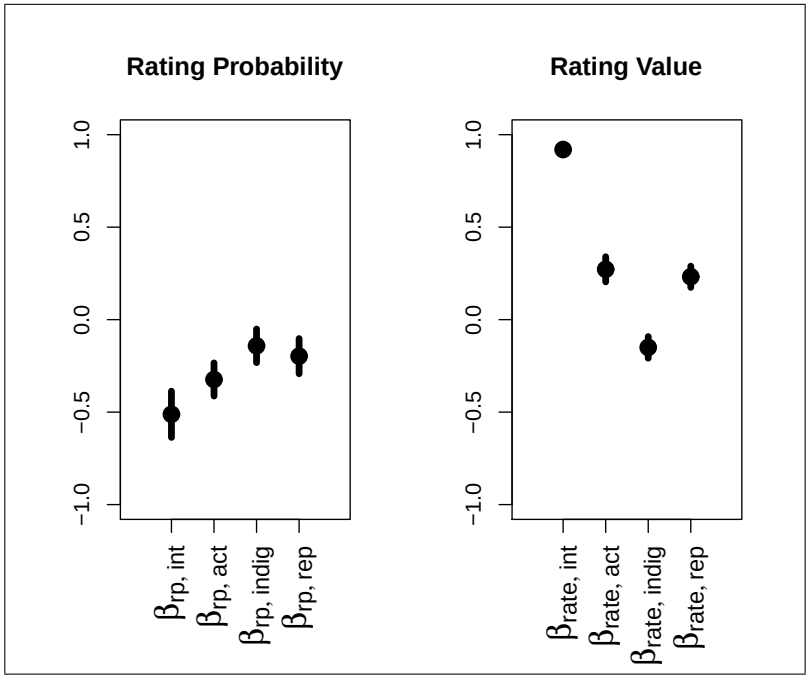


Figure 12. Results from Experiment 5.1, showing 95% HDIs for the posterior distributions of beta weights for the probability of rating as well as the rating value. Note that the rating probability plots are the beta weights at the mean predictor value; the slope changes across the range of all predictors due to the quadratic component on each predictor. Note though, that this change is effectively zero for reputation, and quite small for indignation and actual contribution. See Figure 13 for an illustration. Finally, note that an analysis that did not include the quadratic component yielded qualitatively similar values to those presented here.

punitive application of ratings. Note that the weight on intended contribution is marginally more negative than the weight on actual contribution (Mean difference $=-0.18$, 95% HDI from -0.35 to -0.02). The weight on indignation is small but also negative. Recall that a negative weight on indignation means that the worse the player does in comparison to the target, the *less* likely a reputation rating is to be applied. This is the only weight that trends against intuitively "negative" behaviors being more associated with applying a rating. This is a small weight, but it may be that participants are more likely to fine a high indignation target, and that fining slightly decreases the desire to provide a rating.

In the analysis of probability of rating, the quadratic trend on intended contribution is

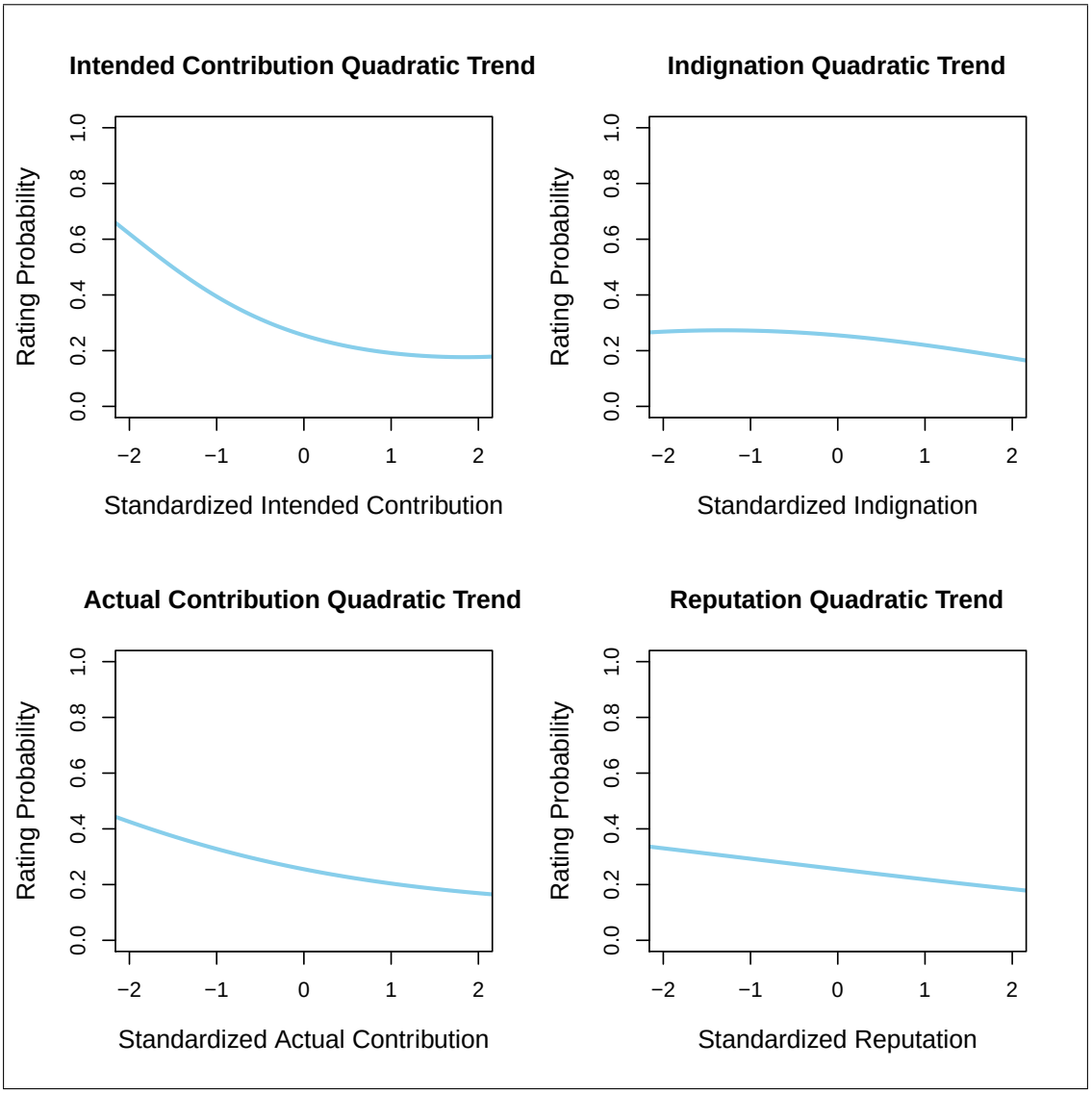


Figure 13. An illustration of the modal quadratic curve predicting the probability of applying a rating as a function of each predictor in Experiment 5.1. Note that the slope at 0 for each predictor is the same as the value plotted in the left half of Figure 12. The predicted probability on the y-axis is assuming all other predictors are at their mean values. Notice that the x-axis is on the standardized scale of of each predictor, meaning that it has a mean of zero and a standard deviation of one. Thus even on a relatively wide scale, the quadratic component does not cause the probability of rating to trend upward for any predictor, even at the high end of the predictor range. However, this quadratic trend does prevent the probability from dropping down below about 0.2 at high levels of intended contribution.

positive and credibly non-zero (Mean=0.13, 95% HDI from 0.06 to 0.20). The quadratic trend for indignation is marginally negative (Mean=-0.06, 95% HDI from -0.11 to -0.004). Neither of the quadratic trends for actual contribution (Mean=0.03, 95% HDI from -0.03 to 0.10) and reputation (Mode=-0.01, 95% HDI from -0.07 to 0.06) are credibly non-zero.

These quadratic trend are illustrated in Figure 13. These plots indicates that the potential U-shaped curve discussed in the description of the model is not really observed for any predictor. That is, the trend does not reverse slope within the observed range of the predictor for any of the four predictors. Even for intended contribution, the probability of rating only begins increasing at very high levels of intended contribution (higher than is even plotted in Figure 13). However, the quadratic component on intended contribution *does* provide additional explanatory information in the form of the small-to-moderate floor in the probability of rating, around 0.2. This floor means that even though ratings are more likely to be used at low intended contribution values, the probability of providing a rating is never extremely low even at the highest intended contributions.

The pattern of regression weights for the value of the ratings shown in the right half of Figure 12 is similar to the pattern of weight for probability of making a rating (though recall that this analysis does not have a quadratic weight on any predictor). Intended contribution dominates, with high intended contribution unsurprisingly associated with a high star rating. Actual contribution and previous reputation are both also moderately positively related to higher star ratings. And indignation is slightly negatively related to rating score, meaning that higher indignation predicts a lower star rating. This makes intuitive sense, and is consistent with an explanation that high indignation tends to lead to fining often in lieu of providing a rating, but if a rating is provided indignation is associated with lower ratings.

There are several take-aways from this analysis. Exclusion and ratings are both heavily

intended-contribution focused. Ratings are most frequently used punitively, but the presence of a positive quadratic trend on intended contribution speaks to the potential of ratings to be utilized as a reward. This potential for use as a reward is further confirmed by looking at the predicted ordinal response for high levels of intended contribution, conditioned on actually applying a rating: at 2 standard deviations above the mean intended contribution (and average levels of the other predictors) the probability of a 5-star response is 50.5%, and the modal probability of a 4-star response is 33.6%.

In the presence of multiple alternative responses that can be used simultaneously, the role of fining appears to change to primarily one of restoring equity between the punisher and the target of the punishment. Given that in actual everyday punishment there are likely multiple response options to a perceived transgression, it is important that future research considers the availability of these alternatives before generalizing to everyday behavior. This is especially salient given that the existing literature overwhelmingly utilizes costly fine as the sole response option.

Chapter 6: Perception of efficacy

In this chapter, we investigate the following research question: does punishment behavior respond to the efficacy of punishment in promoting short-term cooperation?

Cooperation is a hallmark of human social behavior, and an essential component of modern human society. Yet cooperation seems difficult to explain from the perspective of self-interested organisms engaging in competition. This is especially true in cases where individual and group interests conflict, as captured by strategic games like the public goods game and the prisoner's dilemma. The empirical observation of wide-scale cooperation, both in the real world and in the lab, demands an explanation. One potential explanation is punishment.

In a set of landmark results, multiple authors demonstrated the efficacy of punishment in promoting cooperation in the context of strategic games (Fehr & Gächter, 2000; Ostrom et al., 1992; Yamagishi, 1986). However, subsequent research has demonstrated that results such as these are not universal and generalizable to every situation and structure. In some situations, punishment is less able to maintain cooperation. Moreover, in some situations cooperation can be maintained by a punishment mechanism, but causes inefficiency (e.g. Fehr & Gächter, 2000). In this context, "efficiency" is usually measured by the average earnings after all costs (including punishment) are deducted. For punishment to be "inefficient" means that the resources spent enforcing cooperation are greater than the benefits of cooperation being enforced. A large body of work has classified the relevant factors influencing the efficacy and efficiency of punishment in strategic games.

There is copious evidence that punishment is useful for maintaining cooperative behavior (e.g., Balliet et al., 2011; Cinyabuguma et al., 2005; Fehr & Gächter, 2000; Maier-Rigaud et al., 2010; Masclet, 2003; Ostrom et al., 1992; Yamagishi, 1986). However, it is less clear whether a specific individual will adjust their future punishments in response

to the efficacy of their past punishments, or if the urge to punish is a fixed response to perceived transgression.

Cushman (2011) argued that punishment should be a fixed response, in that the likelihood that the punishment will change future behavior should be disregarded by the punisher when deciding to apply punishment. If punishers reduced the magnitude or probability of their punishments when the punishment did not affect the behavior of the punished individual, then persistent transgressors would defeat punishment. Consequently, punishment could not evolve as a mechanism for encouraging cooperation. This argument was borne out by evolutionary simulations (Cushman & Macindoe, 2009). Because cooperation has in fact flourished in real populations, it must be (the argument goes) that punishment evolved to be a fixed response.

Experiment 6.1: Comparing punishment of responsive and non-responsive contributors

Despite the work by Cushman and Macindoe (2009), to our knowledge, responsiveness to efficacy of punishment has not been directly tested. In order to do so, the automated players in Experiment 3.1 were given two types of contribution patterns. The first type we call *punishment-responsive contributors*, who started with relatively high contributions (near 8 points), and reduced their intended contribution by 2 points per round unless a fine was applied, in which case they increased their intended contribution by 2 points. The second contributor type we call *unresponsive contributors*, who started with relatively low contributions (near 3 points) and maintained this low intended contribution consistently, regardless of fines. If efficacy matters to choice of punishment, then unresponsive contributors will be excluded more often and fined less often than responsive contributors, all else being equal.

Predictor Values			Δ Excluding Prob. $P(Exclude)_{\text{responsive}} - P(Exclude)_{\text{unresponsive}}$	Δ Fining Prob. $P(Fine\|\neg Exclude)_{\text{responsive}} - P(Fine\|\neg Exclude)_{\text{unresponsive}}$
Intended	Actual	Indignation	95% HDI	95% HDI
1	2	5	-0.34 to -0.06	0.04 to 0.36
4	1	9	-0.31 to -0.04	0.02 to 0.32
2	2	3	-0.25 to -0.03	0.05 to 0.37
1	1	7	-0.37 to -0.04	0.04 to 0.32
1	2	8	-0.40 to -0.09	0.02 to 0.31

Table 1

Selected predictor values and the corresponding 95% HDIs on the difference between the responsive contributors and unresponsive contributors in probability of exclusion and probability of fine. A positive difference means the probability is higher for responsive contributors than for unresponsive contributors.

Methods. Experiment 6.1 concerns a separate, simultaneous, aspect of the experiment described in Chapter 3. Thus the methods described in regarding Experiment 3.1 apply here. Recall that participants are allowed to apply fines or ostracism or neither.

Results. We performed two separate conditional logistic regressions; one on all the punishment choices in which the target was a responsive contributor, and one on all the punishment choices in which the target was an unresponsive contributor. Because the presence of two contributor types had to be learned by participants, we included only the final 20 rounds in the analysis.

We predict that responsive contributors will be more likely to be fined but less likely to be excluded than unresponsive contributors, *given equal values of the predictors: intended contribution, actual contribution, and indignation*. This hypothesis is agnostic about the relative weights of the predictors, as it only concerns the relative propensities to apply a fine and to apply exclusion across the two types of contributor. To assess this prediction, we took all of the actually occurring combinations of predictor values (intended contribution, actual contribution, and indignation) and computed the propensities to apply exclusion and fine predicted by the two regressions. There were 20,640 such predictor combina-

tions. At each step of the MCMC chain we use the parameter estimates at that step, along with the values of the predictors, to compute posterior predicted probability of exclusion, $P(Exclude)$, and probability of fine given there was not exclusion, $P(Fine|\neg Exclude)$. We then compute the difference of the predicted $P(Exclude)$ and $P(Fine|\neg Exclude)$ values across the two contributor types.[1]

Table 1 displays some predictor sets selected to illustrate the differences between responsive contributors and unresponsive contributors. Consider, for example, the bottom row of Table 1, which indicates a player for whom the intended contribution was 1, the actual contribution was 2, and the indignation was 8. The regression analyses reveal that the probability of excluding that player was about 25 percentage points less if that player was a responsive contributor than if that player was an unresponsive contributor. The 95% HDI on the difference extends from -0.40 to -0.09 (as shown in the table). The probability of fining that player was about about 17 percentage points more if that player was a responsive contributor than if that player was a unresponsive contributor. The 95% HDI on the difference extended from $+0.02$ to $+0.31$ (as shown in the table).

In this set of analyses, we utilize a ROPE (Region of Practical Equivalence) of $\pm.02$. This means that in order for us to consider two probabilities to be credibly different, the 95% HDI on their difference must be entirely less than $-.02$ or greater than $.02$ (see Kruschke, 2018, for a detailed discussion). Using this criterion, we found that for 26% of all the predictor sets, the responsive contributors were credibly less likely to be excluded than unresponsive contributors. No predictor sets showed a credible difference in the opposite direction. Furthermore, 99% of the predictor sets had a mean difference favoring exclusion of unresponsive contributors. In 15% of the predictor sets, responsive contributors were credibly more likely to be fined than unresponsive contributors, and no predictor sets

[1]The average effective sample size (ESS) of the MCMC chain was 8,710 for the estimate of the difference in $P(Exclude)$ across conditions and 5,614 for the estimate of the difference in $P(Fine|\neg Exclude)$ across conditions.

showed a credible opposite trend. 51% of the predictor sets had a mean difference favoring fining of responsive contributors.

These results suggest that participants are sensitive to the efficacy of their punishments, because they punish responsive contributors differently than unresponsive contributors. Participants were making punishment choices not just as an automatic response to freeloading, but were taking into account the potential benefit that could be expected from applying different types of punishment.

Chapter 7: Differences across populations

Experiment 4.2 demonstrated that the emphasis on actual contribution observed in many of the experiments presented here may not be present in all populations. In particular, the population sampled via Amazon Mechanical Turk appears to exhibit a high degree of emphasis on intended contribution regardless of punishment type or experimental manipulation. To our knowledge, no direct comparison of emphasis on actual contribution in punishment behavior has been performed across disparate populations. In this chapter we directly test this difference by comparing Experiment 2.1 to a replication performed on Amazon Mechanical Turk, Experiment 7.1.

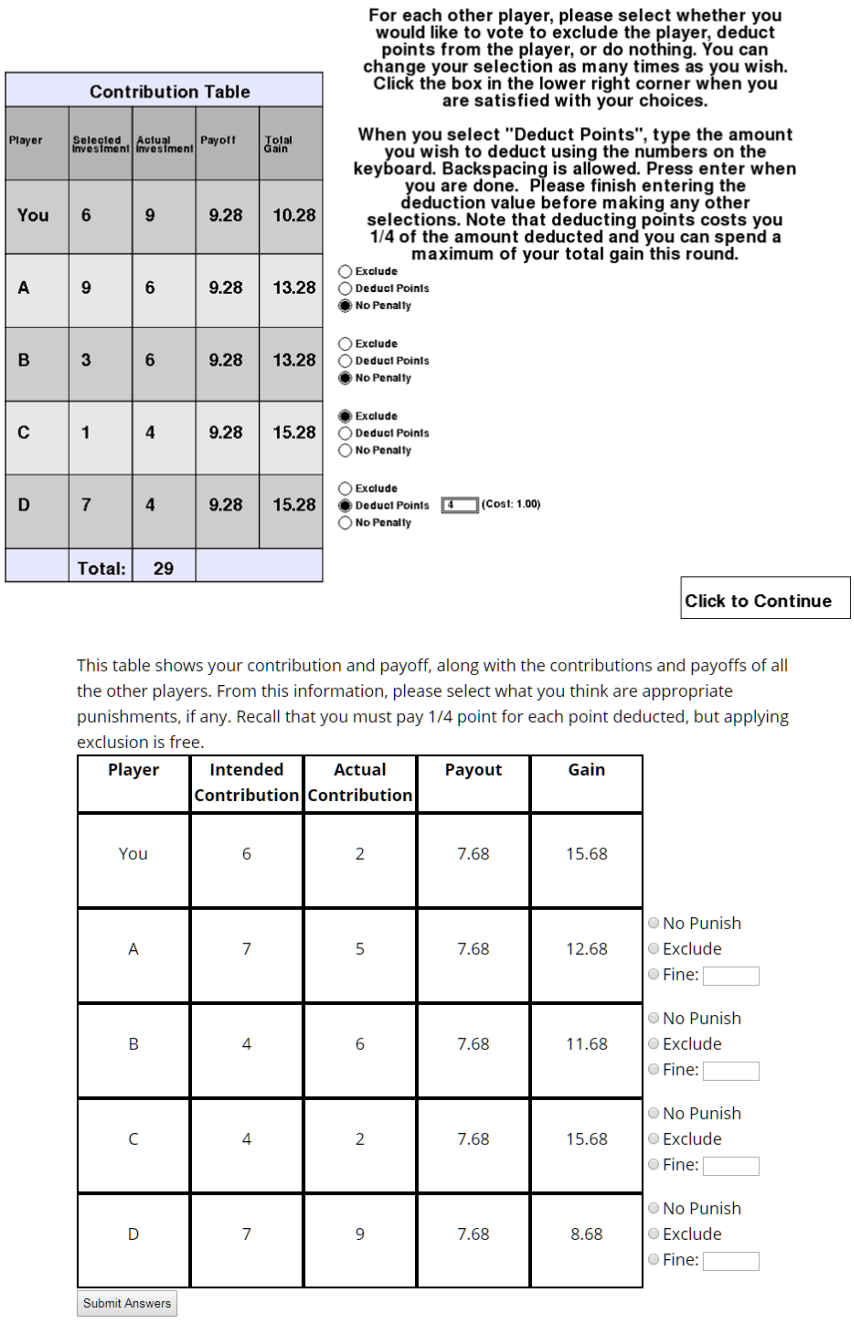
Comparing Experiments 2.1 and 7.1: Two identical experiments in different populations

In this paired set of experiments, we directly investigate if the pattern of emphasis on actual contribution in fining occurs in an online Amazon Mechanical Turk population by directly comparing Experiment 2.1 with a recreation in a web based platform, Experiment 7.1.

Methods. Experiment 2.1 is as described in Chapter 2. Experiment 7.1 is a direct replication in a sample of 146 participants recruited on Amazon Mechanical Turk (AMT). The details of the experimental design for both Experiments 2.1 and 7.1 are presented in Chapter 2. Figure 14 shows a direct comparison of the two interfaces. As in previous online experiments presented here, participants were paid $1.20 for their participation.

These experiments were identical in design to Experiment 2.1: There were two phases of 22 rounds, the first phase had no automated player punishment, and participants had the option to fine or exclude other players, but not both.

Results. We analyzed both sets of results using the same conditional logistic regression model used in previous experiments. The Phase 2 results are summarized in Figure 15.



Contribution Table

Player	Selected Investment	Actual Investment	Payoff	Total Gain
You	6	9	9.28	10.28
A	9	6	9.28	13.28
B	3	6	9.28	13.28
C	1	4	9.28	15.28
D	7	4	9.28	15.28
	Total:	29		

For each other player, please select whether you would like to vote to exclude the player, deduct points from the player, or do nothing. You can change your selection as many times as you wish. Click the box in the lower right corner when you are satisfied with your choices.

When you select "Deduct Points", type the amount you wish to deduct using the numbers on the keyboard. Backspacing is allowed. Press enter when you are done. Please finish entering the deduction value before making any other selections. Note that deducting points costs you 1/4 of the amount deducted and you can spend a maximum of your total gain this round.

Exclude / Deduct Points / No Penalty

Exclude / Deduct Points / No Penalty

Exclude / Deduct Points / No Penalty

Exclude / Deduct Points 4 (Cost: 1.00) / No Penalty

Click to Continue

This table shows your contribution and payoff, along with the contributions and payoffs of all the other players. From this information, please select what you think are appropriate punishments, if any. Recall that you must pay 1/4 point for each point deducted, but applying exclusion is free.

Player	Intended Contribution	Actual Contribution	Payout	Gain	
You	6	2	7.68	15.68	
A	7	5	7.68	12.68	No Punish / Exclude / Fine:
B	4	6	7.68	11.68	No Punish / Exclude / Fine:
C	4	2	7.68	15.68	No Punish / Exclude / Fine:
D	7	9	7.68	8.68	No Punish / Exclude / Fine:

Submit Answers

Figure 14. A comparison of the Experiment 2.1, the university undergraduate sample (above), and Experiment 7.1, the AMT sample (below), punishment choice interfaces. Instructional differences primarily reflect differences in the interface: Experiment 2.1 was administered in a unique environment with input restrictions participants were unfamiliar with, whereas Experiment 7.1 was administered in a web-format without any unfamiliar restrictions.

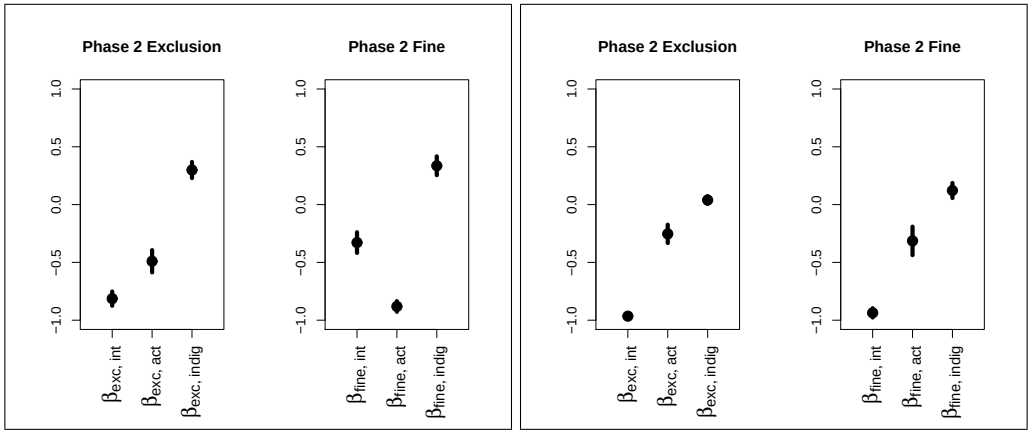


Figure 15. A comparison of the Phase 2 parameter estimates from Experiment 2.1, the university undergraduate sample (left), and Experiment 7.1 the AMT sample (right). As in previous figures, the vertical black bars indicate the 95% highest density interval (HDI) which contains the most credible 95% of the values, with the point indicating the mean. For purposes of space, we show only the results from Phase 2 of both Experiments. Notice how intended contribution weights are much greater in the Experiment 7.1 plot for both exclusion and for fines. Moreover, although this emphasis on intended contribution is slightly higher for exclusion than for fining, the emphasis on intended contribution is so large in both cases that this trend is not credibly non-zero.

The differences across populations are stark. The results from Experiment 7.1 have a high degree of emphasis on intended contribution across both exclusion and fines. In previous work, including Experiment 2.1, outcome-emphasis in fining was a very consistent pattern. However, with only a change in sampling population, this pattern reverses. There does seem to be a slight movement towards outcomes-based decisions in fining, but this change is not credibly different from zero (the HDI on the difference in actual contribution weights across intention and fine is -0.11 to 0.23 in favor of the actual contribution weight for fining). This demonstrates a large departure from the existing published literature that suggests inflexible, outcome-focused punishment decisions (Cushman, 2011; Cushman et al., 2009; Cushman & Macindoe, 2009).

There are several potential reasons why the AMT results differ so greatly from the IU subject pool results. Due to interface differences, the instructions were not completely

identical across the two experiments (see Figure 14 for an illustration of such differences). However, these trivial differences in instructions seem unlikely to cause such a marked difference in behavior. The demographics of Amazon Mechanical Turk are older and more diverse than the IU subject pool (see Paolacci & Chandler, 2014, for a discussion of AMT demographics). This could be a significant driver of the differences. AMT participants are also participating primarily to receive payment, as opposed to course credit. This focus on payment may decrease the emotional salience of the fictitious in-game "points" which did not correlate with payment received, which may mean punitive emotions associated with outcomes and inequality may be reduced. Finally, AMT workers do communicate with one another, including developing communities for the purpose of providing recommendations for particular AMT jobs and otherwise sharing AMT experiences. This "community culture" may lead AMT workers to see the other players not as competitors, but as cooperation partners from the same community. Cooperation partners may be primarily evaluated by their good faith efforts (i.e., intent) rather than the outcomes that occurred.

Discussion

This dissertation presented work with several novel features. We introduced the trembling-hand PGG paradigm, which was essential for testing several hypotheses regarding punishment behavior. In this discussion we briefly discuss the major results presented here, as well as situate the results in the existing literature.

Intention and Outcome

Recall that previous research (Cushman et al., 2009) has demonstrated outcome emphasis in the domain of punishment decision making. Outcome emphasis is an extreme version of outcome bias that occurs when the actual outcome is weighted *more strongly* than the agent's intended outcome. In the context of our PGG, we have referred to this as "emphasis on actual contribution" to avoid confusion with an individual player's total gain.

We have replicated this puzzling phenomenon in some of our experiments. However, we have provided evidence that there is much need for nuance in interpreting previous work on outcome-emphasis in punishment. We observed emphasis on actual contribution in some, but far from all, contexts. In fact, the conditions in which emphasis on actual contribution was observed were quite specific. The pattern was only observed in the case of fines, and even then only in the university undergraduate sample. In all other responses, and in all responses in the Amazon Mechanical Turk (AMT) sample, emphasis on actual contribution was not observed. Moreover, we observed the opposite pattern, emphasis on intended contribution, in a large variety of contexts. In particular, emphasis on intended contribution was observed even when the only change was collecting data in an AMT sample as opposed to a university subject pool, likely resulting from differences in the salience of the in-game points and a culture of community present in this subject pool.

Revisiting automation

In the introduction we summarized literature that suggests that the use of automated players is unlikely to strongly alter the behavior of participants (Barclay, 2006; Fogg & Nass, 1997; Gerstenberg & Lagnado, 2010; Kiesler et al., 1996; Nass et al., 1996; Nass & Moon, 2000; Suri & Watts, 2011).

Our results are consistent with the hypothesis that participants were reacting to the automated players as if they were human. If the participants were treating the automated players merely as unfeeling computer-generated numbers that should be handled in whatever way maximizes personal points, then it is difficult to explain why players should mimic the punishment behavior of the automated players in Experiment 3.1, or why players should administer any costly fines or ratings at all in the experiments where these were available.

Nevertheless, it would be valuable to pursue trembling-hand PGGs with groups of human players. These could include direct replication of the experiments here with human confederates, in which all but only one player is a naive participant. We would tentatively expect results like those we reported here, though we have seen sampling differences yield quite different behaviors, as described in Chapter 7. Another follow-up could involve all naive participants, with the goal being to investigate the contribution and punishment norms that spontaneously arise.

Inequity Aversion

A phenomenon closely related to outcome bias is inequity aversion, wherein punishment is used to enforce equality of outcome (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Raihani & McAuliffe, 2012) or more generally as a response to inequity (Cook & Hegtvedt, 1983; Yamagishi et al., 2009). If individuals are not attempting to punish accidental transgressors and are instead attempting to enforce fairness, outcome bias would be reducible to inequity aversion. Because this motivation depends entirely on the actual

outcome, under some circumstances, inequity aversion could produce behavior that would be classified as outcome emphasis. To test the possibility that outcome emphasis can be reduced to inequity aversion, the die experiment summarized above (Cushman et al., 2009) included a condition in which choosers had *no* control over their allocation to the other player, and allocations were explicitly random. The receiving players still punished "self-ish" allocations, but the effect of outcome was less than in the trembling-hand condition. Therefore, inequity aversion alone is unlikely to be a complete explanation of outcome bias.

In our experiments, we captured a type of personal inequity aversion using our indig-nation predictor. This predictor represents "personal" inequity aversion in that it does not extend to the entire group. In all of our analyses, the regression weight on indignation was non-zero for all punishment types studied. Consistent with previous work, this suggests the some form of inequity aversion plays a role in all punishment decisions in a PGG, and this role is relatively insensitive to the factors manipulated in the experiments presented here. This indignation weight was especially pronounced in the case of fining when multiple non-exclusive punishments were available. This makes sense given that fines are uniquely able to restore equity, unlike reputation ratings and ostracism. This raises the question of what role indignation plays in the case of ostracism and reputation rating, neither of which can restore equity in a direct sense. One potential explanation for this persistent role of in-dignation is that inequity captures attention and causes individuals to immediately consider whether punishment is necessary, after which other information (such as the intended and actual contribution) attenuates or exacerbates the initial impulse to punish.

The only potential exception to the pattern above is in the probability of applying repu-tation ratings. Indignation had the expected effect on the rating applied, if one was applied: high indignation was associated with a low rating value. However, in the case of the prob-ability of rating, indignation plays a reversed role in that higher indignation is associated

with a lower probability of rating. But this effect is small, and applying a rating is not strictly punitive, so this is not inconsistent with the idea that inequity captures attention and further processing of all available information occurs to decide the proper course of action.

Efficacy in Changing Behavior

One purpose of punishment is to promote cooperation. In the context of individual decision making this raises a significant question: Will individuals stop punishing if the punishment stops fulfilling its purpose (i.e., the punishment fails to encourage cooperative behavior)? There are competing intuitions as to the answer. On the one hand, it seems that punishers should be able to evaluate whether a punished individual has changed behavior, and also that they would not wish to keep spending resources punishing a repeat offender. On the other hand, it seems emotionally negative to "reward" a repeat offender with a lack of punishment, and the emotional motivation behind punishment may not be sensitive to more practical concerns like efficacy (e.g., Xiao & Houser, 2005; Yamagishi et al., 2009).

In Experiment 6.1, we found evidence that individuals are sensitive to the efficacy of their punishments when deciding what punishments to apply. Recall that Cushman (2011) argued that punishment behavior that is extinguishable by inefficacy is easily exploitable by cheaters who ignore punishment, and these arguments were supported by results from evolutionary simulations by Cushman and Macindoe (2009). While this claim seems at odds with Experiment 6.1, it might be attributable to the assumption of Cushman and Macindoe (2009) that the only choice available to punishers is to apply a punishment or not. On the contrary, in the present experiment and many real world interactions, there are a range of potential punishment responses. In the arguments presented by Cushman, the choice is to punish or not to punish, and furthermore carrying out this punishment has a direct cost to the punisher. Baumard (2010, 2011) argues that costly punishment is not representative of human punishment behavior, at least in the conditions representative of the environment

early humans would have faced. Instead, Baumard argues that exclusionary punishments that might be as simple as a refusal of repeat interactions are far more common. If we consider that individuals in a punishment scenario have a trinary choice of applying a fine (or other material punishment), applying an exclusionary punishment, or doing nothing, it becomes less clear that being sensitive to efficacy would be selected against. In such a situation, a punisher could apply fine punishments to non-cooperators who they believe could be responsive, apply an exclusionary punishment to those who they believe will be unresponsive, and do nothing to cooperators. It is not obvious that such a situation would provide a selection pressure towards being insensitive to punishment. On the contrary, it may select for "conditional" non-cooperators who gain some of the benefits of free-riding and then in the future gain all the benefits of cooperation. We did not investigate efficacy in contexts that also allowed public ratings, but the addition of a third reasonable punishment to behavior that is deemed undesirable could even further support the usefulness of flexibility.

Norms of Punishment

Experiment 3.1 demonstrated that experimentally-manipulated social norms of punishment can potentially change the punishment behavior of participants. Moreover, one potential explanation for the sample differences observed in Chapter 7 is a difference in pre-existing norms exogenous to the experimental design. Both of these speak to the importance of norms in punishment behavior, which further supports the general trend of flexibility in punishment behavior.

Efficiency and the Cost of Punishment

We have discussed the role of punishment in promoting cooperation, and the efficacy of punishment for promoting cooperation is well-validated (Fehr & Gächter, 2000). However,

some researchers have investigated whether this cooperative effect is worth the cost of punishment, both in terms of the cost of applying punishments and in terms of the benefits lost via the application of punishment. This topic has been studied from several angles.

Perhaps the most obvious consideration when studying punishment efficiency is the cost of a fine relative to the amount of the fine (e.g., Carpenter, 2007; Nikiforakis & Normann, 2008). The higher this amount, the less efficient punishment is (Nikiforakis & Normann, 2008) and the less the punishment is utilized (Carpenter, 2007). This has some intersection with the experiments described in Chapter 4 that systematically varied whether a given punishment had a cost. We found weak evidence that the costly punishment conditions saw less punishment: punishment occurred in 26% of opportunities in the free conditions, versus 17% of opportunities in the costly conditions, but this difference was not enough to yield a credibly non-zero difference. We did not directly assess efficiency, and the existence of a meaningful "efficiency" for an exclusionary punishment is not immediately obvious, but we did not find major differences in the behavior of participants across the costly and cost-free conditions.

The potential for counter-punishment is another issue to consider when assessing punishment efficiency. Counter-punishment refers to the application of punishment not as a response to the primary behavior (in our case, the contribution to the public good) but in response to a punishment received in a previous round. Counter-punishment could both decrease efficiency by decreasing the payout of the "counter-punished" individual, as well as by discouraging the use of punishment to promote cooperation. These possibilities have been investigated in the context of public goods games (Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008). In these experiments, the presence of counter-punishment both decreased the likelihood of punishment being utilized at all, as well as decreased the overall efficiency of the punishment scheme.

We did not investigate counter-punishment directly, and in experiments where auto-

mated players punished, they were not programmed to counter-punish. However, counter-punishment intersects with the punishment types discussed insofar as it is not possible in the case of ostracism: once someone is ostracized they will be unable to counter-punish the person or people responsible for ostracizing them. This is yet another example of the importance of utilizing multiple punishment types.

In summary, the existing literature relating the cost and efficiency of punishment should be supplemented with investigation of punishments that can reasonably have low or zero cost, yet still maintain cooperation, like ostracism or reputation ratings. Notably, these punishments also do not affect the gain of the target player, further increasing the monetary efficiency of punishment schemes using these punishment types.

Punishment type

As we have frequently stated, the existing literature on punishment is greatly focused on costly fine (e.g. Fehr & Gächter, 2000; Fehr & Schmidt, 1999; Ostrom et al., 1992). We utilized fines, both free and costly, in nearly all of the experiments presented here. We have found that fines are the only punishment type that tends to be focused on outcomes rather than intentions, although this pattern was not always observed. In Experiment 5.1, with three independent responses available, fining was especially associated with indignation. We also found that this punishment was used over ostracism when the punished individuals improved their behavior in response to fining.

In addition to fines, this dissertation has addressed two other real-world punishments, ostracism and reputation damage. Previous research on ostracism has demonstrated that it can motivate cooperation in public goods games (Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010; Masclet, 2003), may be more representative of human everyday punishment than costly fines (Baumard, 2010, 2011; Baumard et al., 2013) and is more frequently observed in non-human animals (Raihani & McAuliffe, 2012; Stevens et al., 2005). In our

experiments, we found ostracism to be very focused on intended contribution, often in direct contrast to fines. This pattern was attenuated, though not completely reversed, when participants were presented with a local norm of actual-contribution-focused ostracism. The intended contribution focus pattern was not found to be sensitive to the cost of ostracizing. We also found ostracism to be utilized over fining when individuals did not improve their behavior in response to fining.

The final punishment type we studied was reputation damage (simultaneously with its reward counterpart, reputation improvement). Reputation is important to daily life (Kurland & Pelled, 2000; Madden & Smith, 2010), especially in the marketplace (Gertsen et al., 2006; Rhee & Valdez, 2009), and previous research has found that reputation is essential to cooperation via indirect reciprocity (Mohtashemi & Mui, 2003; Nowak, 2006; Sommerfeld et al., 2008; Wang et al., 2012). We found reputation ratings to be primarily used punitively, but with a sort of "floor" that ensures that reputation ratings are used moderately often even for good behavior. And the ratings themselves are overwhelmingly determined by intended contribution. Pre-existing reputation was used as a signal to apply future punishments, though it did not have as large an influence as current behavior.

In general, we have found that the less commonly studied alternative punishments investigated here (reputation influence and ostracism) tend to be similar in terms of the behaviors they target. In fact, we have provided evidence that fining may be a unique punishment among punishment types. No other punishment method was observed to exhibit emphasis on actual contribution, under any of the experimental manipulations or conditions we investigated. This included a norm-manipulation that was explicitly intended to produce emphasis on actual contribution, but did not sufficiently alter behavior to produce this pattern. All of this supports a greater emphasis on alternative punishments in the existing punishment literature. Moreover, our results highlight the importance of context in the punishment behavior, both in terms of the features of the experimental apparatus as well

as the properties of the sample being studied. These unexpected contextual differences support the importance of replication, both direct replications in different populations and variations in experimental design.

Finally, the punishments we have studied here are far from exhaustive. There are many punishments (or even more generally, responses to undesired behavior) utilized in everyday interactions that we did not attempt to replicate in our experimental setup, including direct communication of discontent, inflicting emotional pain, inflicting physical pain, and likely others. It is possible that these examples fall into clear "types" similar to those studied here, but it is almost certain that some have unique features important to understanding the breadth of human punishment behavior.

Conclusions

This dissertation covered many topics regarding punishment behavior in our novel paradigm, the trembling hand PGG. In particular, this paradigm allowed us to assess the different roles of intention and outcome in punishment behavior, a novel possibility in the PGG paradigm. We also re-implemented this paradigm in a web-based format to allow data collection from any internet-connected computer. We investigated multiple punishments over and above the standard costly fine, presented side-by-side and alone, finding these alternative punishments to vary importantly from the typically used costly fines. Our novel automated player design allowed complete experimental control and the testing of novel hypotheses that required this level of control. We used customized Bayesian models to assess the hypotheses of interest, allowing our analysis methods to match the structure of the experiment and the goals of the analysis. We found that alternative punishments are used frequently, used intelligently, are importantly unique in their application, and thus their inclusion in punishment research is essential going forward.

More than any singular conclusion, we have found evidence for flexibility in pun-

ishment behavior. In contrast to theories of punishment that describe it as unilaterally outcome-emphasizing, unresponsive to efficacy, or otherwise singular in character, we have found copious evidence that punishment behavior resists any simple characterization. The tremendous variability in punishment behavior based on punishment cost and type, punishment efficacy, different populations, experimentally manipulated local norms, and other factors all support this conclusion. These results taken together suggest that punishment is not solely a simple instinctual response to a perceived transgression (though that may be an important component). Contrary to such an account, punishment is a flexible and nuanced tool individuals utilize to navigate and manipulate their complex and changing social environment.

References

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin*, *137*(4), 594.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(September 2005), 325–344. doi: 10.1016/j.evolhumbehav.2006.01.003

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of personality and social psychology*, *54*, 569–579. doi: 10.1037/0022-3514.54.4.569

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, *9*(2), 171–192. doi: 10.1007/s11299-010-0079-9

Baumard, N. (2011). Punishment is not a group adaptation. *Mind & Society*, *10*(1), 1–26. doi: 10.1007/s11299-010-0080-3

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: the evolution of fairness by partner choice. *The Behavioral and brain sciences*, *36*(1), 59–78. doi: 10.1017/S0140525X11002202

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review*, *90*(1), 166–193.

Book, A. S. (1999). Shame on you: An analysis of modern shame punishment as an alternative to incarceration. *Wm. & Mary L. Rev.*, *40*(2), 653 – 686.

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization*, *62*(4), 522–542.

Carpenter, J. P., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics*, *12*(3), 272–288. doi: 10.1007/s10683-009-9214-z

Carpenter, J. P., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters*, *86*(3), 393–398. doi: 10.1016/j.econlet.2004.08.007

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*(8), 1421–1435. doi: 10.1016/j.jpubeco.2004.05.011

Cook, K. S., & Hegtvedt, K. A. (1983). Distributive justice, equity, and equality. *Annual review of sociology*, *9*(1983), 217–241. doi: 10.1146/annurev.so.09.080183.001245

Cushman, F. (2011). Moral Emotions from the Frog's Eye View. *Emotion Review*, *3*(3), 261–263. doi: 10.1177/1754073911402398

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS one*, *4*(8), e6699. doi: 10.1371/journal.pone.0006699

Cushman, F., & Macindoe, O. (2009). The Coevolution of Punishment and Prosociality Among Learning Agents. In *Proceedings of the 31st annual conference of the cognitive science society.*

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*(1), 145–167. doi: 10.1007/s00199-007-0212-0

Denwood, M. J. (2013). runjags: An R package providing interface utilities, distributed computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. doi: 10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817–868.

Flynn, C. (1994). Regional differences in attitudes toward corporal punishment. *Journal of Marriage and the Family*, *56*(2), 314–324.

Fogg, B. J., & Nass, C. (1997). How Users Reciprocate to Computers : An experiment that demonstrates behavior change. *CHI'97 extended abstracts on Human factors in computing systems*, 331–332. doi: 10.1145/1120212.1120419

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian*

Data Analysis (Third ed.). Chapman and Hall/CRC Texts in Statistical Science.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–71. doi: 10.1016/j.cognition.2009.12.011

Gertsen, F. H., van Riel, C. B., & Berens, G. (2006). Avoiding reputation damage in financial restatements. *Long Range Planning*, *39*(4), 429–456.

Henrich, J., Boyd, R., Bowles, S., Camerer, C. F., Fehr, E., Gintis, H., ... Tracer, D. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *The Behavioral and brain sciences*, *28*, 795–815; discussion 815–855. doi: 10.1017/S0140525X05000142

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–7. doi: 10.1126/science.1153808

Jacobs, D., & Carmichael, J. (2002). The political sociology of the death penalty: A pooled time-series analysis. *American Sociological Review*, *67*, 109–131.

Jones, J. M. (2013). U.S. death penalty support lowest in more than 40 years. *Gallup Politics*. Retrieved from `http://www.gallup.com/poll/165626/death-penalty-support-lowest-years.aspx`

Kahan, D. M. (1996). What do alternative sanctions mean? *U. Chi. L. Rev.*, *63*(2), 591–653.

Kahan, D. M. (2006). What's really wrong with shaming sanctions. *Yale Law School Legal Scholarship Repository Faculty Scholarship Series*, *102*.

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of personality and social psychology*, *70*(1), 47–65. doi: 10.1037/0022-3514.70.1.47

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2013). Posterior predictive checks can and should be Bayesian: comment on

Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'. *The British Journal of Mathematical and Statistical Psychology*, *66*(1), 45–56. doi: 10.1111/j.2044-8317.2012 .02063.x

Kruschke, J. K. (2015). *Doing Bayesian data analysis, second edition: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012, sep). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. doi: 10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. doi: 10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206.

Kurland, N. B., & Pelled, L. H. (2000). Passing the word: Toward a model of gossip and power in the workplace. *Academy of management review*, *25*(2), 428–438.

Lansford, J. E., & Dodge, K. A. (2008). Cultural norms for adult corporal punishment of children and societal rates of endorsement and use of violence. *Parenting: Science and Practice*, *8*(3), 1–11. doi: 10.1080/15295190802204843.Cultural

Liddell, T. M., & Kruschke, J. K. (2014). Ostracism and fines in a public goods game with accidental contributions: the importance of punishment type. *Judgment and Decision Making*, *9*(6), 523–547.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley and sons.

Luce, R. D. (2008). Luce's choice axiom. *Scholarpedia*, *3*(12), 8077.

Madden, M., & Smith, A. (2010). Reputation management and social media. *Pew Internet & American Life Project*.

Maier-Rigaud, F. P., Martinsson, P., & Staffiero, G. (2010). Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior & Organization*, *73*(3), 387–395. doi: 10.1016/j.jebo.2009.11.001

Masclet, D. (2003). Ostracism in work teams: a public good experiment. *International Journal of Manpower*, *24*(7), 867–887. doi: 10.1108/01437720310502177

Mohtashemi, M., & Mui, L. (2003). Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism. *Journal of Theoretical Biology*, *223*(4), 523–531.

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*, 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81. doi: 10.1111/0022-4537.00153

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves?âŸĘ. *Journal of Public Economics*, *92*(1-2), 91–112. doi: 10.1016/j.jpubeco.2007.04.008

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, *11*(4), 358–369. doi: 10.1007/s10683-007-9171-3

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560. doi: 10.1126/science.1133755

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Wiley.

Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, *457*(7225), 79.

Ostrom, E., Walker, J. M., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *The American Political Science Review*, *86*(2), 404–417.

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing.*

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*(July), 18–21. doi: 10.1098/rsbl.2012.0470

Rhee, M., & Valdez, M. E. (2009). Contextual factors surrounding reputation damage with potential implications for reputation repair. *Academy of Management Review*, *34*(1), 146–168.

Sasaki, T., & Uchida, S. (2013). The evolution of cooperation by social exclusion. *Proceedings. Biological sciences / The Royal Society*, *280*(1752), 20122498. doi: 10.1098/rspb.2012.2498

Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1650), 2529–2536.

Stevens, J. R., Cushman, F., & Hauser, M. D. (2005). Evolving the Psychological Mechanisms for Cooperation. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 499–518. doi: 10.1146/annurev.ecolsys.36.113004.083814

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE*, *6*(3). doi: 10.1371/journal.pone.0016836

Walker, J., & Halloran, M. (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics*, *7*(3), 235–247.

Wang, Z., Wang, L., Yin, Z.-Y., & Xia, C.-Y. (2012). Inferring reputation promotes the evolution of cooperation in spatial social dilemma games. *PloS one*, *7*(7), e40218.

Whitman, J. Q. (1998). What is wrong with inflicting shame sanctions? *Yale Law School Legal Scholarship Repository Faculty Scholarship Series*, *655*.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7398–401. doi:

10.1073/pnas.0502399102

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116. doi: 10.1037/0022-3514.51.1.110

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Science*, *106*(28), 11520–11523. doi: 10.1073/pnas.0900636106

Zeisel, H., & Gallup, A. (1989). Death penalty sentiment in the United States. *Journal of Quantitative Criminology*, *5*(3), 285–296.

Appendices: Notes on Data Analysis

In these appendices, we provide detailed descriptions of all the models utilized in data analysis, labeled by experiment, and organized such that similar analyses are discussed in sequence.

Appendix 1: Details of the model for Experiment 1.1

The application of a fine by player i to player j is denoted $y^{[i,j]}$ and is an integer from the set $\{1, 2\}$ where 1 denotes no penalty was applied, and 2 denotes that a fine was applied. It should be noted that this analysis models the propensity to apply a fine, not the *amount* of the fine. (Please see Appendix 4 for an analysis that incorporates the amount of fine with analogous results.) The model describes the action, $y^{[i,j]}$, as a random draw from a categorical distribution (i.e., multinomial distribution with $N = 1$) with category probabilities $\pi_{none}^{[i,j]}$ and $\pi_{fine}^{[i,j]}$ (with the constraint that $\pi_{none}^{[i,j]} = 1 - \pi_{fine}^{[i,j]}$), which is denoted

$$y^{[i,j]} \sim \text{cat}\left(\pi_{none}^{[i,j]}, \pi_{fine}^{[i,j]}\right) \tag{1}$$

where the symbol "$\sim$" is read "is distributed as," and where cat indicates a categorical distribution.

The model uses three predictors. One predictor is the intended contribution by player j, denoted $x_{int}^{[j]}$. Intuitively, as the intended contribution increases, the probability of punishing the player should decrease. The second predictor is the player's actual contribution, denoted $x_{act}^{[j]}$. Intuitively, as the actual contribution increases, the probability of punishing the player should decrease.

The third predictor is what we call the "indignation" of subject i toward player j, which is the net gain of player j minus the net gain of subject i, denoted $x_{indig}^{[i,j]}$. Intuitively, as indignation increases, the probability of punishment might increase. Indignation is in-

cluded as a predictor to reflect inequity aversion, and because the net gain was explicitly displayed on the screen along with intended and actual contributions. As described in the introduction, previous experiments have demonstrated inequity aversion in punishment. By including a separate predictor for inequity aversion, the independent influences of actual and intended contributions can be better assayed.

We used a standard logistic regression model for describing each individual player. For each subject i, we compute a weighted combination of the predictors for the underlying tendency to apply a fine to a player j, denoted $\lambda_{fine}^{[i,j]}$:

$$\begin{aligned}\lambda_{fine}^{[i,j]} = \beta_0^{[i]} & \\ & + \beta_{int}^{[i]} \left(x_{int}^{[j]} - \overline{x}_{int} \right) \\ & + \beta_{act}^{[i]} \left(x_{act}^{[j]} - \overline{x}_{act} \right) \\ & + \beta_{indig}^{[i]} \left(x_{indig}^{[i,j]} - \overline{x}_{indig} \right) \end{aligned} \tag{2}$$

Equation 2 shows that the predictors were mean-centered by subtracting their overall means across all trials and players. This mean centering makes the intercept, $\beta_0^{[i]}$, better interpretable as baseline behavior at the mean of the predictors, and makes shrinkage from the hierarchical model (to be described below) apply at the mean instead of at zero.

The underlying tendency to apply a fine is converted to a probability via the conventional logistic function:

$$\phi_{fine}^{[i,j]} = 1 \Big/ \left(1 + \exp(-\lambda_{fine}^{[i,j]}) \right) \tag{3}$$

To produce the final probability of applying a fine, the model accounts for "oops" errors by mixing the probability of Equation 3 with a random-choice probability of 1/2, using a

mixing coefficient α:

$$\pi_{fine}^{[i,j]} = \alpha\,(1/2) + (1-\alpha)\left(\phi_{fine}^{[i,j]}\right) \tag{4}$$

Effectively, Equation 4 makes the logistic function have asymptotes at $\alpha/2$ and $1 - \alpha/2$ instead of at 0 and 1. It is worth noting that estimates of the guessing rate α were quite small, with typical values not exceeding 0.009 in any analysis. Nevertheless, including a non-zero α is important to account for rare outlying responses that could otherwise force the regression coefficients to be artificially small in magnitude.

We use a hierarchical model in which individual $\beta^{[i]}$ coefficients are assumed to come from higher-level distributions that describe group-level tendencies. Each individual's coefficients are assumed to be t-distributed across the group:

$$\begin{aligned}
\beta_0^{[i]} &\sim \mathrm{t}(\mu_0, \tau_0, \nu=5) \\
\beta_{int}^{[i]} &\sim \mathrm{t}(\mu_{int}, \tau_{int}, \nu=5) \\
\beta_{act}^{[i]} &\sim \mathrm{t}(\mu_{act}, \tau_{act}, \nu=5) \\
\beta_{indig}^{[i]} &\sim \mathrm{t}(\mu_{indig}, \tau_{indig}, \nu=5)
\end{aligned} \tag{5}$$

where τ is the precision (reciprocal of squared scale) of the t-distribution, and where ν is the normality of the distribution, often referred to as the degrees-of-freedom parameter. Preliminary analyses indicated considerable unsystematic outliers in the individual-level predictor coefficients. Therefore we choose the relatively low value of 5 for ν to allow the group-level coefficients to be robust against individual outliers. The use of t distributions to accommodate outliers is routine in statistical modeling (e.g., Kruschke, 2013).

The primary focus of the analysis is on the group-level means of the regression coefficients in Equation 5. The estimate of μ_{int}, for example, is the group-level mean value for

$\beta_{int}^{[i]}$, which is the weight placed on the intended contribution for applying a fine. It should be noted that in all the figures we plot a normalized reparameterization of the regression weights according to the following formula, where $pred$ is a placeholder for int, act, or $indig$:

$$\text{Normalized}(\mu_{pred}) = \frac{\mu_{pred}}{\sqrt{\mu_{int}^2 + \mu_{act}^2 + \mu_{indig}^2}} \tag{6}$$

The normalization across predictors is reasonable because the scales of the three predictors are the same: monetary points. The normalized regression weights represent the values of the raw regression weights relative to one another. This allows easier comparison across regression weights, and in later experiments across conditions. We refer to the normalized group-level μ_{pred} parameters as "beta weights" because they denote the typical values of the coefficients in Equation 2.

The hierarchical structure of the model rationally imposes shrinkage on the individual estimates. The estimate of each $\beta_{pred}^{[i]}$ is influenced by subject i's responses and by the estimates of the higher-level μ_{pred} and τ_{pred} parameters. The higher-level parameters are influenced by data from all subjects, hence each individual's estimate is a compromise between the individual's data and the typical group data. Hierarchical models are an especially useful way to estimate group-level tendencies, without assuming that all individuals have identical behavior, and without assuming that all individuals are mutually uninformative (e.g., Gelman & Hill, 2007; Kruschke, 2015). It is important to note that due to the mean centering of the predictors (see Eqn. 2) the intercept expresses baseline behavior at the mean values of the predictors and thus shrinkage applies to the mean-centered intercepts and slopes. This makes the shrinkage more meaningful than applying it to intercepts located arbitrarily at zero monetary points, which for the actual and intended contribution predictors essentially never occurred in the experiment.

We establish vague, noncommittal prior distributions for the means and precisions of the group distributions:

$$\mu_{pred} \sim \text{normal}(0, 1e-10)$$

$$\tau_{pred} \sim \text{gamma}(1.10512, 0.010512)$$

where the shape and rate constants in the gamma distribution give it a mode of 10 and a standard deviation of 100. These broad prior distributions imply that the prior has minimal influence on the posterior distribution. The α parameter also had a noncommittal prior, $\alpha \sim \text{uniform}(0, .1)$.

Appendix 2: Details of the model for Experiments 2.1, 3.1, and 6.1

These experiments add an additional punishment option in the form of exclusion to the basic framework of Experiment 1.1. This punishment option was mutually exclusive with the application of a fine. This structure requires a model that allows for three possible outcomes.

The model we utilized is a conditional logistic regression, where the first probability (the probability of exclusion) is modeled as in Appendix 1, and one outcome is modeled conditional on the first outcome not occurring: the probability of applying a fine, given that no exclusion was applied.

As discussed in the main text, a traditional analysis for n-ary choice data is multinomial logistic regression, which models the probabilities of all choices without conditionalizing on any one of them. However, we believe out data violates the assumption of the independence of irrelevant alternatives, which is a required assumption for this type of model (Luce, 1959, 2008).

As in Appendix 1, we denote the punishment applied by subject i to player j as $y^{[i,j]}$, which is now an integer from the set $\{1, 2, 3\}$ where 1 indicates no punishment, 2 indicates a fine, and 3 indicates exclusion. The model assumes that $y^{[i,j]}$ can be described as a random draw from a categorical distribution:

$$y^{[i,j]} \sim \text{cat}\left(\pi_{none}^{[i,j]}, \pi_{fine}^{[i,j]}, \pi_{exc}^{[i,j]}\right) \tag{7}$$

We use the same predictors and logistic function as in the analysis of Appendix 1. Thus, we treat the underlying tendency for subject i to apply exclusion to player j, $\lambda_{exc}^{[i,j]}$ as a weighted

combination of the predictors:

$$\begin{aligned}\lambda_{exc}^{[i,j]} = \beta_{exc,0}^{[i]} \\ &+ \beta_{exc,int}^{[i]}\left(x_{int}^{[j]} - \overline{x}_{int}\right) \\ &+ \beta_{exc,act}^{[i]}\left(x_{act}^{[j]} - \overline{x}_{act}\right) \\ &+ \beta_{exc,indig}^{[i]}\left(x_{indig}^{[i,j]} - \overline{x}_{indig}\right)\end{aligned} \tag{8}$$

Furthermore, $\lambda_{fine}^{[i,j]}$ is the underlying tendency for subject i to apply a fine to player j, given that subject i did not exclude player j:

$$\begin{aligned}\lambda_{fine}^{[i,j]} = \beta_{fine,0}^{[i]} \\ &+ \beta_{fine,int}^{[i]}\left(x_{int}^{[j]} - \overline{x}_{int}\right) \\ &+ \beta_{fine,act}^{[i]}\left(x_{act}^{[j]} - \overline{x}_{act}\right) \\ &+ \beta_{fine,indig}^{[i]}\left(x_{indig}^{[i,j]} - \overline{x}_{indig}\right)\end{aligned} \tag{9}$$

These underlying tendencies are converted to choice probabilities as follows:

$$\begin{aligned}\phi_{exc}^{[i,j]} &= 1 \Big/ \left(1 + \exp(-\lambda_{exc}^{[i,j]})\right) \\ \phi_{fine}^{[i,j]} &= \left[1 \Big/ \left(1 + \exp(-\lambda_{fine}^{[i,j]})\right)\right]\left(1 - \phi_{exc}^{[i,j]}\right) \\ \phi_{none}^{[i,j]} &= 1 - (\phi_{fine}^{[i,j]} + \phi_{exc}^{[i,j]})\end{aligned} \tag{10}$$

The conversion to choice probabilities in Equation 10 is what makes the model *conditional* logistic regression, because the probability of fining is the logistic of the fining tendency multiplied by the probability of not excluding. It should also be noted that due to the way $\phi_{exc}^{[i,j]}$ and $\phi_{fine}^{[i,j]}$ are defined, $\phi_{none}^{[i,j]}$ cannot be less than zero. The regression coefficients of Equations 8 and 9 are estimated using the conditional probabilities of Equation 10.

As in the analysis in Appendix‘1, the logistic probabilities of Equation 10 are mixed with random choices (1/3) to accommodate occasional off-task responses or “oops” errors:

$$\pi_{action}^{[i,j]} = \alpha\,(1/3) + (1-\alpha)\left(\phi_{action}^{[i,j]}\right) \tag{11}$$

The probabilities of Equation 11 are used to model the trinary choices in Equation 7.

In summary, for each individual we have two sets of beta weights, one set describing the propensity to apply exclusion, and the other set describing the propensity to apply a fine given that exclusion was not applied.

We again use a hierarchical model in which individual beta coefficients are assumed to come from higher-level distributions that describe group-level tendencies. Each individual’s coefficient is assumed to be t distributed across the group:

$$\beta_{pen,pred}^{[i]} \sim \mathrm{t}(\mu_{pen,pred}, \tau_{pen,pred}, \nu = 5) \tag{12}$$

where the subscript pen stands in for either of the two possible penalties (exclude or fine) and the subscript $pred$ stands in for any of the three predictors (intended contribution, actual contribution, or indignation). As in Appendix 1, the primary focus of the analysis is on the group-level means $\mu_{pen,pred}$ in Equation 12.

For the Bayesian estimation, we use the noncommittal prior distributions that were used for the analysis of Appendix 1. And, like the analysis of Appendix 1, we use MCMC techniques to generate 20,000 representative credible values from the joint posterior distribution on the 2,825 parameters in each phase. Unless otherwise noted, the effective sample size for all results reported in the article was at least 10,000.

Appendix 3: Details of the model for Experiments 4.1, 4.2, and 7.1

The model for these experiments was a very slight variation on the model described in Appendix 2. First, in the model for these experiments, we no longer fixed the value of the normality parameter ν. Instead, it was estimated with the following prior:

$$\nu \sim \exp(1/30) \tag{13}$$

This elaboration allows the level of non-normality in the distribution of participants to be estimated as opposed to being pre-specified, but tend to be similar to the pre-specified value used in Appendix 2

Appendix 4: Details of the models for Experiment 5.1

The analysis of Experiment 5.1 has two primary components. Recall that this experiment had non-exclusive responses in the form of ratings, fines, and exclusion. The first component applies to the probability of applying exclusion and fine independently. Each of these is modeled separately using a logistic regression similar to the one described in Appendix 1, with two elaborations. The first is that we again estimate the normality parameter ν as in Appendix 3. The second is that we standardize the predictors prior to the analysis. This is to allow direct comparability between predictor weights, as this analysis has a fourth predictor, pre-existing reputation, that is not on the same "points" scale as the other three predictors. Other than these small changes, the logistic portion is largely identical to the model in Appendix 1.

The novel portion of the analysis comes from the analysis of ratings. This analysis has two sub-components, the probability of applying a rating, and the model of the ratings values. The probability of applying a rating is again a logistic model much like the model of exclusion and fine, with one elaboration. We expected that the probability of applying a rating may not be monotonic with each of the predictors. For instance, a low intended contribution may be associated with a high probability of rating (likely a low rating), and a high intended contribution may be associated with a high probability of rating (likely a high rating), with middling intended contribution having low probability of rating. This potential "U-shape" is not possible using only linear predictors. To accommodate this we include quadratic predictors for each of the four predictors. This means that the tendency for player i to apply a rating to player j, denoted $\lambda_{rate}^{[i,j]}$, is a linear combination of the four

predictors, plus a quadratic component (denoted β_{predQ}) for each predictor:

$$\begin{aligned}\lambda_{rate}^{[i,j]} =&\beta_0^{[i]} + \beta_{int}^{[i]}x_{int}^{[j]} + \beta_{act}^{[i]}x_{act}^{[j]} + \beta_{indig}^{[i]}x_{indig}^{[i,j]} + \beta_{rep}^{[i]}x_{rep}^{[j]} + \\ &\beta_{intQ}^{[i]}(x_{int}^{[j]})^2 + \beta_{actQ}^{[i]}(x_{act}^{[j]})^2 + \beta_{indigQ}^{[i]}(x_{indig}^{[j]})^2 + \beta_{repQ}^{[i]}(x_{rep}^{[j]})^2\end{aligned} \tag{14}$$

Note that in this equation $x_{pred}^{[j]}$ represents the predictor value on a *standardized* scale, for reasons discussed earlier in this appendix.

The rating value portion of this analysis consists of an ordered-probit regression. In this analysis, the rating of player j by player i is denoted $y^{[i,j]}$ and is an integer from the set $\{1, 2, 3, 4, 5\}$. This value is modeled as a categorical distribution:

$$y^{[i,j]} \sim \text{cat}\left(\pi_1^{[i,j]}, ..., \pi_5^{[i,j]}\right) \tag{15}$$

Note that this analysis does not include "no response" as it is conditional on the response occurring. The response occurring is modeled using the logistic model described previously.

The probability of a given rating k, denoted $\pi_k^{[i,j]}$, is a function of a thresholded-cumulative normal distribution, with mean μ, standard deviation σ and a set of thresholds θ:

$$\pi_k^{[i,j]} = \Phi\left(\frac{\theta_k - \mu^{[i,j]}}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \mu^{[i,j]}}{\sigma}\right) \tag{16}$$

where $\Phi()$ is the standardized cumulative normal function. This equation says that the probability of rating k is the area under the normal curve between threshold θ_{k-1} and threshold θ_k. For the first level (i.e., $k = 1$) the threshold θ_{k-1} is negative infinity, and for the highest level (i.e., $k = 5$) the threshold θ_5 is positive infinity. In other words, the probability of rating 1 is determine by the left tail of the normal distribution below θ_1 and the probability

of rating 5 is the right tail of the normal distribution above θ_4.

The value μ, representing the central tendency of the underlying normal distribution, is a linear combination of the predictors, much like the λ parameters of previous models:

$$\mu^{[i,j]} = \beta_0^{[i]} + \beta_{int}^{[i]} x_{int}^{[j]} + \beta_{act}^{[i]} x_{act}^{[j]} + \beta_{indig}^{[i]} x_{indig}^{[i,j]} + \beta_{rep}^{[i]} x_{rep}^{[j]} \tag{17}$$

Note again that $x_{pred}^{[j]}$ represents the predictor value on a standardized scale. The θ thresholds are also estimated:

$$\theta_k \sim \text{normal}(k + 0.5, 2) \tag{18}$$

with the exception of the first and fourth thresholds which are fixed at $\theta_1 = 1.5$ and $\theta_4 = 4.5$

The remainder of the parameters have non-committal, vague priors as in previous analyses. The values of the predictors are also hierarchical in structure, with each participant being modeled by their own individual beta-weights, centered on a group estimate in a t distribution, also as described in the previous models.

Torrin M. Liddell

Education

Ph.D. Psychology and Cognitive Science, Indiana University. November 2018.

B.S. Psychology and Philosophy, Michigan State University. High Honors. 2011.

Publications and Submitted Manuscripts

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328-348.

Casey, K., Cullen, A., Landis, L., Liddell, T. M., Mason, D., Schumm, J., Tandy, K. *Indiana Task Force on Public Defense Report.* Available at https://www.in.gov/publicdefender/2333.htm

Breithaupt, F., Li, B., Liddell, T.M., Brower, E., & Whaley, S. (2018). Fact vs. Affect in the Telephone Game: An Examination of Surprise in Story Retelling. *Frontiers in Psychology, 9*, 2210.

Eyink, J.R., Motz, B.A., Heltzel, G., Liddell, T.M. (2018). The role of "Fit" in social norm interventions: The case of self-regulated studying behaviors. *In review.*

Kruschke, J. K. & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178-206.

Kruschke, J. K. & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*, 155-177.

Liddell, T. M., & Kruschke, J. K. (2014). Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. *Judgment and Decision Making, 9*(6), 523-547.

Breithaupt, F., Gardner, K. M., Kruschke, J. K., Liddell, T. M., and Zorowitz, S. (2013). The disappearance of moral choice in serially reproduced narratives. In: M. A. Finlayson, B. Fisseni, B. Lowe, and J. C. Meister (Eds.), Workshop on Computational Models of Narrative, pp. 3642. Germany: Dagstuhl Publishing.

Presentations

Eyink, J., Motz, B., Hetzel, G., & Liddell, T. M. (2017). November 10th. Online self-directed learning activities, and the norms that influence them. IU Online Conference 2017. Presented by Ben Motz.

Eyink, J., Motz, B., Hetzel, G., & Liddell, T. M. (2017). October 27th. Online self-directed learning activities, and the norms that influence them. Colloquium, Indiana University, Bloomington. Presented by Ben Motz.

Liddell, T. M., & Kruschke, J. K. (2016). May 27th. Analyzing ordinal data: Support for Bayesian Ordinal Models. Poster presented at The 28th Association for Psychological Science Conference.

Liddell, T. M., & Kruschke, J. K. (2016). May 28th. Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. Poster presented at The 28th Association for Psychological Science Conference.

Liddell, T. M., & Kruschke, J. K. (2015). November 22nd. Analyzing ordinal data: Support for a Bayesian approach. Poster presented at The Psychonomics Society Conference 2015.

Liddell, T. M., & Kruschke, J. K. (2015). November 21st. Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. Poster presented at The Society for Judgment and Decision Making Conference 2015.

Liddell, T. M., & DeDeo, S. (2015). June 25. Inequality, community, and common knowledge: The effects of meta-knowledge in cooperation on networks. Colloquium, The Santa Fe Institute.

Liddell, T. M., & Kruschke, J. K. (2014). January 24. Towards a More Flexible View of Punishment: Results from a Trembling-Hand Public Goods Game. Colloquium, Indiana University, Bloomington.

Liddell, T. M., & Kruschke, J. K. (2013). June 14. Towards a More Flexible View of Punishment: Results from a Trembling-Hand Public Goods Game. Invited talk presented by John Kruschke at Universität Basel, Switzerland.

Liddell, T. M., & Kruschke, J. K. (2013). April 3. Towards a more flexible view of punishment: Results from a trembling-hand public goods game. Colloquium, Indiana University, Bloomington.

Liddell, T. M., & Pleskac, T. J. (2011). April 8. Subadditivity in probability judgments and working memory. Poster presented at Michigan State University Undergraduate Research and Arts Forum.

Teaching

Indiana University

Statistical Techniques (PSY-K300). Fall 2016. Associate Instructor. Pilot course for an experimental undergraduate statistics course that included a large Bayesian component.

Advanced Statistics in Psychology (PSY-P553, Prof. John K. Kruschke). Fall 2014 and Fall 2015. Lab Instructor.

Introduction to Bayesian Data Analysis (PSY-P533, Prof. John K. Kruschke). Spring 2015, Spring 2016, and Spring 2017. Teaching Assistant.

Methods of Experimental Psychology (PSY-P211, Ben Motz). Spring 2014. Lab Instructor.

Statistical Techniques (PSY-K300, Cynthia Patton). Summer 2016. Teaching Assistant.

Statistical Techniques (PSY-K300, Prof. S. Lee Guth). Summer 2013, Fall 2013, Summer 2014, and Summer 2015. Teaching Assistant.

Cognitive Psychology (PSY-P335, Thomas Gruenenfelder). Spring 2013. Teaching Assistant.

Special Topics in Psychology: Science of Moral Judgment (PSY-P457, PSY-P657, Prof. John K. Kruschke). Fall 2012. Teaching Assistant.

Michigan State University

Appointed tutor, Department of Philosophy. 2010.

Honors and Awards

College of Arts and Sciences Graduate Student Travel Award, Indiana University. Fall 2015.

Commendation on Ph.D. Qualifying Exam. 2014.

Indiana University Psychological and Brain Sciences Departmental Fellowship. Fall 2011 to Spring 2016.

Michigan State University Study Abroad Scholarship. Sophia University (Tokyo, Japan). Summer 2010.

Michigan Competitive Scholarship, Michigan State University. Fall 2010 to Spring 2011.

Psi Chi Psychology Honors Fraternity Membership, Michigan State University Fall 2009 to Present.

Honors College Membership, Michigan State University. Fall 2008 to Spring 2011.

Deans List, Michigan State University. Fall 2007 to Spring 2011

Michigan Promise Scholarship, Michigan State University. Fall 2007 to Spring 2009.