When Punishment Doesn't Pay: Cold Glow and Decisions to Punish

Aurélie Ouss *Harvard University*
Alexander Peysakhovich *Harvard University*

Abstract

Economic theories of punishment focus on determining the levels that provide maximal social material payoffs. In other words, these theories treat punishment as a public good. Several parameters are key to calculating optimal levels of punishment: total social costs, total social benefits, and the probability that offenders are apprehended. However, levels of punishment are often determined by aggregating individual decisions. Research in behavioral economics, psychology, and neuroscience shows that individuals appear to treat punishment as a private good (cold glow). This means that individual choices may not respond appropriately to the social parameters. We present a simple theory and show in a series of experiments that individually chosen punishment levels can be predictably too high or too low relative to those that maximize social material welfare. Our findings highlight the importance of the psychology of punishment for understanding social outcomes and for designing social institutions.

1. Introduction

The criminal justice system is an expensive segment of modern society, but it has an important instrumental role: it helps ensure cooperation and social order. Since Becker (1968), there has been a large interest in the economics of crime and punishment (see Levitt and Miles [2007] for an empirically minded review). The Beckerian framework focuses on levels of punishment that yield optimal outcomes, where the marginal (material) costs of punishment equal the marginal

We would like to thank Phillipe Aghion, Yochai Benkler, Tom Cunningham, Ed Glaeser, Roland Fryer, Oliver Hart, Drew Fudenberg, Louis Kaplow, Lawrence Katz, David Laibson, David Rand, Steve Raphael, Al Roth, Steven Shavell, Bruce Western, and the seminar participants at the Harvard Labor Lunch for helpful comments. Part of this research was funded by a grant from the Lab for Economic Applications and Policy at Harvard University. Ouss is a Terence M. Considine Fellow in Law and Economics at Harvard Law School and acknowledges support from the Considine Family Foundation and the Harvard Law School's John M. Olin Center for Law, Economics, and Business. Peysakhovich thanks the John Templeton Foundation for financial support.

[*Journal of Law and Economics,* vol. 58 (August 2015)]

(material) benefits of decreased crime.[1] Thus, the Beckerian approach can be used as a normative theory of how to set punishment levels.

In many cases, however, levels of punishment in society are determined by aggregating individual decisions. For example, voters change laws (directly or via representatives), juries composed of civilians deliver verdicts, and sometimes groups or individuals mete out social punishments themselves. In this paper we use experimental methods to ask two questions: First, do individual decisions about punishment respond to parameters that are important for setting Beckerian optimal punishments? Second, when levels of punishment are chosen by individuals, are socially optimal outcomes reached?

Researchers in the behavioral sciences have recently become interested in understanding human punishment behavior. In lab settings, where punishment is formally defined as the willingness to take actions that reduce the payoffs of others, a large portion of individuals are willing to incur costs to punish those who act in inappropriate ways (Ostrom, Walker, and Gardner 1992; Fehr and Gachter 2000; Peysakhovich, Nowak, and Rand 2014), even when they have no personal stake (Fehr and Fischbacher 2004) or no possible strategic motive (Fudenberg and Pathak 2010). These findings suggest that individuals punish because they directly derive utility from reducing the payoffs of those who violate norms of cooperation.[2] Studies in social neuroscience support this theory: activity in the brain's reward areas during costless punishment can be used to predict punishment behavior in costly punishment situations (De Quervain et al. 2004), and reward activity is not visible when cooperative players are punished (Singer et al. 2006). We refer to this broadly defined set of individual motivations as cold glow, in reference to warm-glow theories of altruism in which individuals receive utility from the act of being cooperative beyond the final social consequences of cooperation (Andreoni 1990). Additional evidence suggests that the proximal mechanism that drives cold glow involves affective considerations: strong negative emotions are engaged when other individuals violate social norms (Xiao and Houser 2005; Fehr and Gächter 2002).[3] Finally, research in moral psychology hints that this motive is very blunt.[4] Taken together, this broad array of evidence

[1] There exist multiple channels by which punishment can increase cooperation: deterrence, specific deterrence, and incapacitation (Shavell 1987) are often mentioned. These economically motivated analyses share a common thread: they each view punishment as a means to ensure social cooperation.

[2] Although there is considerable evidence that what constitutes a violation of cooperation norms appears to vary by society (Henrich et al. 2010; Herrmann, Thöni, and Gächter 2008) and can be manipulated in the lab (Peysakhovich and Rand, forthcoming).

[3] Recent research in neuroscience (Sanfey et al. 2003; Knoch et al. 2006) suggests that both affective and controlled processes are important for punishment behavior, and Buckholtz et al. (2008) find that controlled processes might matter more in determining criminal responsibility, while affective processes are more engaged when choosing magnitude of sanctions.

[4] Cushman et al. (2009) asked individuals to play a modified dictator game in which the dictator chose between dice, with each die yielding different probabilities of fair or selfish allocations. After the die was rolled, recipients were allowed to punish or reward the dictator. The authors find that outcomes predict punishment or reward behavior by the recipients, while intentions (choice of die) have a smaller effect. In a similar vein, Coffman (2011) finds that when defection takes place via an intermediary, punishment behaviors are reduced.

raises the question of whether individual punishment decisions will be reactive to changes in more abstract parameters, such as the probability of apprehension or total social costs and benefits. If not, aggregates of individual punishment behaviors might not result in optimally deterring levels of punishment and may indeed lead to inefficient social outcomes.

We note that our focus is not on pinning down the exact mechanisms driving punishment behavior. Rather, we treat cold-glow motivations as a first-order approximation and focus on asking how these well-documented facets of individual psychology can interact with institutions, modeled as available punishment technologies,[5] creating inefficiencies at the aggregate level. More specifically, we ask whether aggregates of individual decisions reach levels of Beckerian optimal deterrence.

We hypothesize that individuals may respond to private costs much more than to social costs, and thus when these costs are externalized, for example, via group-funded punishment, individuals may demand a much higher level of punishment than is consistent with optimal deterrence. In addition, if individuals are driven by more blunt "just deserts" motivations, they may ignore the role of probability of apprehension. In this case, environments in which individuals are rarely caught may have aggregate punishment levels that are too low to deter expected-utility-maximizing offenders. Finally, optimally deterring punishments take into account total levels of sanction, but cold-glow punishers' decisions may not be crowded out by other punishers' choices.[6] As our main contribution, we explore cold-glow punishment in a series of lab experiments that are designed not only to look for individual motives but also, importantly, to relate individual decisions to aggregate outcomes. To look at the effects of cost sharing, probability of apprehension, and crowding out, we present three experiments in which people can punish a particular norm violation: taking from a third party. Our experimental designs allow for transparent calculation of levels of punishment that would reach the optimal deterrence benchmark: not only can we ask whether individual behavior responds to particular parameters, but we can also explore whether aggregate outcomes are optimal in some sense.

Our first experiment examines how punishment choices are affected by cost structures: we vary whether the costs of implementing punishment are borne by the individuals making the choice or by the group. The punishment used in this experiment is excluding norm breakers from the game: when they are excluded, they can neither make money nor take from other players. Our experi-

[5] There is a substantial literature that looks at the differential effectiveness of different punishment mechanisms in various games (Ostrom, Walker, and Gardner 1992; Xiao and Houser 2011; Houser et al. 2008; Andreoni and Gee 2012; Sutter, Haigner, and Kocher 2010; Nikiforakis 2008; Casari and Luini 2009; Balafoutas, Grechenig, and Nikiforakis 2014). In our analysis, however, we set a mechanism and vary parameters of the environment as opposed to setting an environment and varying the mechanism. The use of findings from psychology to design punishment mechanisms robust to changes in parameters is an important topic for future research.

[6] We choose these three facets because they affect policy-relevant behaviors. We survey existing evidence of potential cold-glow effects in the field in Section 5.

mental design is such that the use of relatively small punishments can result in social goals consistent with maximizing overall cooperation; yet when costs are not fully internalized, players overpunish. Our second experiment investigates the role of probability of apprehension in punishment choices. A player can take from a third party, and we experimentally vary the probability with which he is caught and punished. We compare ex ante punishment choices and taking behavior across conditions. Choices of penalty are not affected by changes in the probability of apprehension, but takers' behavior is. This leads to a different kind of inefficient punishment: levels are too low to deter socially destructive behavior.

Our final experiment looks at whether our cold-glow terminology is apt. The theory of warm glow posits that individuals gain private benefits from the act of contributing to a public good and not from the overall amount. We ask whether individuals gain private benefits from overall levels of punishments imposed on norm breakers or whether these psychic benefits come from their own contributions to the punishment. In our study, two individuals make punishment decisions in sequence. We look at whether the second decision maker's punishment decreases with the punishment of the first individual and find that, on average, no crowd out occurs.

We note that some of these effects have been demonstrated in second-party contexts, when the punisher's material welfare had been affected by the offense (for example, Anderson and Putterman [2006] and Nikiforakis and Normann [2008] demonstrate a demand curve for punishment, while Casari and Luini [2012] and Duersch and Müller [2010] discuss imperfect crowding out). In these experiments, unlike in ours, a motive of personal revenge always exists when punishing a norm breaker. Our results complement the literature on peer punishment by showing that many results continue to hold even in situations involving third-party punishment. Thus, our experiments also indirectly shed some light on the question of whether second- and third-party punishments are instantiated via similar psychological mechanisms. Our experimental design is also more representative of settings of interest to legal scholars, as conviction and sentencing are more akin to third-party than to second-party punishment.

2. Beckerian versus Cold-Glow Punishers

2.1. A Simple Reduced-Form Punishment Model

Traditional economic theories of crime examine the case of rational criminals. Here we derive our predictions in a simple reduced-form model, and in the online appendix we present a more detailed game-theoretic derivation of these results for a single actor.

Consider a single-shot scenario in which a continuum of individuals can choose to engage in an action that is personally beneficial (they gain benefit b) but socially costly. If individuals choose to take this action, they are caught with probability p and receive a punishment of size l. Suppose that b varies across individuals for various exogenous reasons. (Although it is an important subject,

we do not discuss the responses of punishers to changes in the distribution of b in the population.) This means that for a given probability of being caught and a punishment level, we can write a demand curve $D(p, l)$, which is the amount of socially inefficient action that occurs and is decreasing in both p and l. We assume that demand is smooth and downward sloping. For simplicity we assume that even at $l = \infty$, there is some wasteful action that is taken because of either trembles or random utility models.[7]

First, let us suppose that a money-maximizing planner chooses the level of punishment l with all other parameters held exogenously fixed. He considers a social loss function $V[D(p, l)]$ (which we assume is convex) and a cost function, which we take as linear for simplicity, cl. This means for a given set of parameters there is an optimal punishment level $l^*_{\text{Becker}}(p, D, c)$ that minimizes the total social costs:

$$V[D(p, l)] + cl.$$

We refer to this as the Beckerian optimum.

Note that the Beckerian optimum has two important comparative statics: first,

$$\frac{\partial l^*_{\text{Becker}}}{\partial c} < 0;$$

that is, the optimal Beckerian punishment decreases in social cost per unit of punishment. This is because the optimum equates the marginal benefit of another unit of punishment (that is, decreases in defection) with the marginal cost. Second, we have

$$\frac{\partial l^*_{\text{Becker}}}{\partial p} < 0.$$

As the probability of being caught decreases, the marginal benefit of a unit of punishment decreases (by the convexity assumption above); thus, levels of punishment decrease.

Now suppose that the planner is an individual who can set l but bears a fraction γ of society's cost (for example, through taxes). The individual puts weight θ on total social welfare (that is, has social preferences). He or she also might derive some direct utility $G(l)$ from the act of punishment.[8]

The individual's maximization problem becomes to choose l to maximize

$$-\gamma cl + G(l) - \theta\{-V[D(p, l)] - (1 - \gamma)cl\}.$$

We refer to the case of $G(\cdot) = 0$ as a Beckerian punisher.

This simple model provides some immediate insights. First, if punishers are Beckerian for any nonzero γ, we obtain punishment levels that are different from

[7] This could be achieved, for example, if all individuals have logistic random utility as in a quantal response equilibrium model (McKelvey and Palfrey 1995) or if they always tremble to an unintended action with small probability.

[8] Note that $G(l)$ need not necessarily be strictly increasing in sanction. For example, a person might want a fair punishment that fits the offense.

the socially optimal level (essentially because punishment here is a public good). On the other hand, if punishers indicate a demand for punishment as a private good, then shifting costs to society may actually increase punishment above the Beckerian optimum. In addition, we also see that if θ is small relative to $G(\cdot)$, then individual decisions will not respond to changes in probability of apprehension as much as they should.

Finally, suppose that punishment is split into two parts, so individuals who are caught first receive a punishment of l_1 and then a punishment of l_2. Note that the Beckerian optimum is applicable to the joint punishment $l = l_1 + l_2$, so optimal choices of l_2 are decreasing in l_1. This means that Beckerian punishers should find crowding-out effects of others' punishments on their own. On the other hand, if punishers care about their own contribution rather than total levels,[9] then crowding out may not occur, which can result in higher levels of punishment than the Beckerian optimum.

These very different responses to parameters lead to very different predictions of the effects of various institutions. Making punishment easier by reducing or shifting the costs or having multiple punishers can result in good outcomes in a world where individuals punish for Beckerian reasons but can result in inefficiently high levels of punishment in a world where individuals choosing punishments are motivated by cold glow. Significantly lowering the probability of apprehension may lead to punishment levels that are too low, and, finally, allowing for multiple punishers can result in total amounts of punishment that could be seen as excessive.

2.2. Experimental Design

We now turn to evaluating our discussion empirically. Our experiments test two types of questions. The first are comparative statics: Do punishment levels respond to social or private costs? Is punishment crowded out? And does another important parameter in the Beckerian model—probability of apprehension—change punishment decisions made by individual punishers?[10] We note that these individual-level questions are about comparative statics rather than levels and so do not require strong assumptions about the form of utility functions.

The second set of questions we seek to address are about mechanism design: are punishments too low or too high relative to Beckerian benchmarks? To answer these questions, we require some assumptions about utility functions, and for simplicity we choose risk neutrality and assume that individuals tremble to unintended actions with probability ε. Without this assumption, punishment levels of infinity lead to infinite deterrence and are weakly preferred to any other

[9] This logic leads to predictions of a lack of crowding in cooperation under the warm-glow theory of public-goods provision (Andreoni 1993; Cornes and Sandler 1994).

[10] In the full model, available in the online appendix, we show that even for ex ante punishments, if decision makers have preferences for punishments that fit the crime, when the probability of apprehension is low, punishers might still shy away from the (high) levels of punishment that would sustain low levels of offending.

Table 1

Summary of Experiments

	Punishment	Probability of Apprehension	Other Punishment	Payment Structure
Experiment 1: responses to costs:				
Social costs	Exclude	1/2	None	Punisher pays
Private costs	Exclude	1/2	None	Group pays
One-round take[a]	Exclude from game	1/2	None	Group pays
Experiment 2: probability of apprehension:				
Low	Take away payoffs	1/3	None	Punisher pays
High	Take away payoffs	9/10	None	Punisher pays
Experiment 3: crowding out				
Response to other punishers' decisions	Take away payoffs	3/4	Other punisher	Punisher pays
Response to computer decision	Take away payoffs	3/4	Random computer punishment	Punisher pays

[a] Only one round of taking plus punishment.

punishment level. While these assumptions are necessarily restrictive, they allow us to make statements about whether aggregate actions lead to socially optimal outcomes and, if not, how far from optimal they are. Table 1 summarizes the experimental designs.

3. Experiment 1: Responses to Costs

In the first experiment, we ask an individual-level question and a group-level question. At the individual level, we test whether costs of punishment accruing to the group rather than to the individual lead to higher demand for punishment. At the social level, the game is set up so that very low levels of punishment are sufficient to deter potential norm breakers. We then ask, will aggregate outcomes be in line with the Beckerian benchmark of optimal deterrence?[11]

3.1. Experimental Design

We ran a series of experiments in which we varied the availability and cost structure of sanctions. In our game, participants gained monetary units (MU) throughout the experiment, which were converted into dollars at a rate of 50 MU per dollar. Players were randomly matched into groups of eight to 12 players. Each group was given a public pot of 70*n* MU (where *n* is the number of players), which was equally split among all members of the group at the end of the game. Each player was also given 30 MU at the beginning of the game.

Participants played 20 rounds (one iteration) of the following game. They were asked to solve a simple math problem, for which they received 4 MU on completion. They were then given the possibility to take. If a player chose to take, she received 2 MU, and another randomly selected player lost 3 MU. Taking is a socially destructive behavior in this case; yet, in the absence of sanctions, it is a dominant strategy. When a player chose to take, she was caught in 50 percent of cases. Our conditions and treatments consist of varying what happens when a player is caught.

In the No Punishment condition, when a player was caught, she received a message informing her that she had been caught, but nothing more happened. In both punishment conditions, when a player was caught, another random player was chosen to be her assigner. The assigner was able to punish players who were caught by excluding them from the game for up to 10 rounds. We elicited punishment using the strategy method: individuals chose a punishment after making their decision of whether to take and seeing whether they were taken from but before they were informed of whether they were caught or if they were someone's assigner. They were then asked to enter the number of penalty rounds that they would assign if they were chosen as an assigner for this round. Individuals could never be chosen as their own assigner, nor did assigners know to whom they as-

[11] There are other potential public-good motivations at play here beyond deterrence, such as incapacitation or specific deterrence. We discuss them in more detail later as well as in the online appendix.

signed penalty rounds. In particular, if they were taken from, there was no additional chance that they would assign a punishment to the player who took from them. In all conditions, only the assigner and the individual to whom penalty rounds were allocated learned about the punishment level.

Each round of exclusion is costly, and we varied the cost structure. In the Private Punishment (hereafter Private) condition, if a player's punishment was chosen, he paid 2 MU from his private funds for each round of punishment he imposed. In the Public Punishment (hereafter Public) condition, if a player's punishment was chosen, each round cost 5 MU from the public pot. This means that in the Public condition, the private share of the cost to a particular punisher was less than 2 MU per round. This experimental design allows us to investigate cost effects in the demand for punishment, thus determining whether the demand for punishment looks like the demand for a public good, as stipulated in most economic models of law enforcement.

As a robustness check, we included one more condition. In the One-Round Take condition, subjects played one round in which they could take and punish (with the public cost structure), followed by 10 rounds in which the option to take was not available. In this case, since subjects could not take in the subsequent rounds, future-oriented motives (incapacitation or deterrence) cannot explain their choice of punishment. This is similar to the design employed in Fudenberg and Pathak (2010), in which individuals played multiple rounds of public-goods games that included sanctions, but the total levels of sanctions chosen were revealed only at the end of the session.

In each experimental session, individuals were first put into a group to play one iteration in the No Punishment condition. After random rematching into new groups, they played either one iteration in the Public condition, one iteration in the Private condition, or three iterations in the One-Round Take condition.[12,13] We implemented this design for several reasons: it allows individuals to gain experience with the experiment in the first stage, and it allows us to look for correlations between individual behavior in the No Punishment condition and later behavior when punishment is available.

Our experimental design is different from other experimental designs that assess the role of nonaltruistic motives for punishment. We vary the cost structure of punishment, which allows us both to discuss the institutional setup of financing sanctions and to investigate the private benefits from punishment, using a basic economics framework. Second, the punishment in the game is not fines, as in prior experiments, but exclusion for a certain number of rounds. This allows us

[12] Participants were not informed about the full structure of the experiment; they were given instructions only for their current condition. They were informed when the One-Round Take condition was the final game in the experiment.

[13] Given lab size constraints, several sessions were conducted for each treatment, but subjects could participate only once in the experiment. They were not informed that later sessions of the same experiment would take place, which made it implausible that players had ulterior deterrence in this experiment in mind.

to include an analysis of incapacitation and therefore contribute to the discussion of different incarceration motives in the economics-of-crime literature.

The experiment was conducted in June and July 2012 at the Harvard Decision Science Laboratory using the software z-Tree (Fischbacher 2007).[14] The participants, recruited using the Decision Science Laboratory's pool of volunteers, were university students (mean age of 21.5, 58 percent female) in the Boston area. We had a total of 91 participants: 39 in the Public, 28 in the Private, and 24 in the One-Round Take conditions.

Participants were given a $10 show-up fee, and their earnings were converted at a rate of 50 MU per dollar. The experiment took between 40 and 50 minutes to complete. Participants earned between $17 and $23. They were informed of the earnings for each condition independently, and their final earnings were privately announced to them at the end of the experiment.

Our main outcome of interest in this series of experiments is the choice of number of rounds of punishment for potential takers who are caught. This is our measure of how large a sanction players are willing to support when facing different cost structures.

3.2. Theories of Punishment

There are three major theories of punishment in the law and economics literature: incapacitation, general deterrence, and specific deterrence. Our experimental design allows us to examine what kind of social benchmark each of these motives sets. We briefly present predictions of these different theories of choices of punishment; a full discussion is in the online appendix.

Incapacitation prevents offending by removing offenders. Shavell (1987) determines the optimal level of punishment for achieving cost-efficient incapacitation. He finds that for incapacitation to be cost-efficient, the cost of incarceration (or, in our experiment, of removing a player for N rounds) must be lower than the cost of an individual's expected harm if not incapacitated. In the Public condition, the cost of incapacitation outweighs its benefits, which makes it an insufficient motive to explain positive punishment levels.

General deterrence is the impact of the threat of future punishment on behaviors. In our setup, players cannot increase general deterrence by setting higher levels of punishment. Only players who are caught learn about other players' punishment choices, and even then they know only their assigner's choice of penalty rounds. General threats cannot be issued.

Specific deterrence is the impact of received sanctions on offenders' future behaviors: receiving larger sanctions could make offenders who are caught less likely to take in future rounds. This motive could be a consideration, and we explore it formally in the online appendix, making different assumptions about takers' behaviors. The main result is that sanctions should be decreasing as the game nears its end, and, regardless of our assumptions about takers' behaviors, in the

[14] Instructions for the experiments are available in the online appendix.

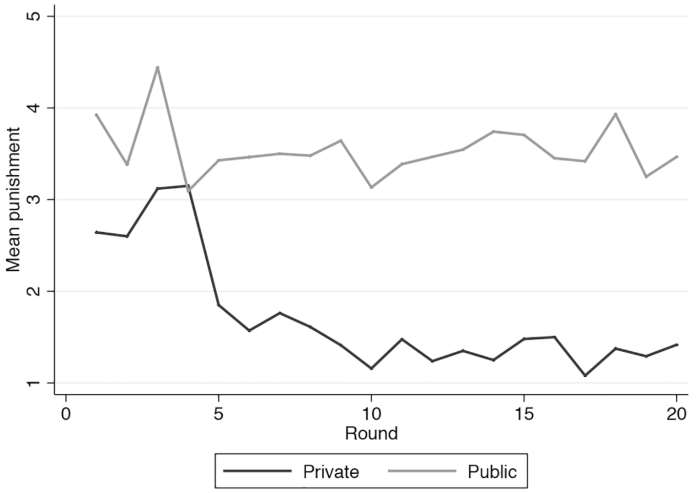

Figure 1. Mean punishment level chosen by round

One-Round Take condition, no sanction can be rationalized by a specific deterrence motive, since taking is possible only in the first round of the condition.

Unlike with prosocial motives, cold glow predicts that punishment levels in the Public condition will be higher than in the Private condition. Private benefits from cold-glow motives will be overconsumed when costs are not fully internalized. In addition, cold glow is the only motivation consistent with any nonzero punishment in the One-Round Take condition.

3.3. Results of Experiment 1

This section compares the Public with the Private condition. We then present additional evidence from the One-Round Take condition as a robustness check.

3.3.1. Punishment Decisions

We first look at punishers' decisions.[15] Figure 1 presents the number of rounds of punishment chosen in the Public and Private conditions.[16] There is a learning effect, but after five rounds, punishment levels in the Private condition drop substantially below those in the Public condition. The average punishment levels

[15] We note for completeness that in the middle of two of the experimental sessions, a bug in the software caused group accounts to unintentionally gain an extra 20–30 MU. No participants reported noticing the gain, participants' behavior appears not to have been affected by the event, and all our results are robust to restricting our analyses to rounds before this occurrence.

[16] As a reminder, all players who were not currently excluded from the game could choose a punishment.

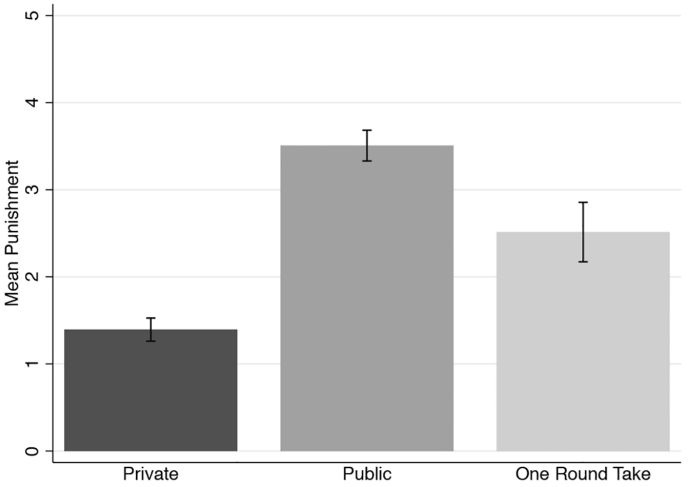

Figure 2. Mean punishment level chosen; rounds >5 only for the Public and Private conditions.

settle to 1.3 rounds in the Private condition and stays at 3.5 rounds in the Public condition.

The fact that punishment levels remain the same over rounds in the Public condition is a first indication that specific deterrence cannot be the only motivation at play: as participants get closer to the end of the game, the size of imposed punishment does not decline. The idea behind specific deterrence is that players who are punished have some period of time during which to apply the lessons that they have learned and no longer take; fewer rounds would be necessary for that as the end of the game gets closer.[17] Furthermore, the average levels of punishment chosen in the Public condition far exceed the levels that would be expected for optimal deterrence or incapacitation.

3.3.2. Robustness Check

To conclusively rule out deterrence or incapacitation as the only motives for punishment, we also consider the One-Round Take condition. Figure 2 shows the average punishment decisions made in all iterations of the One-Round Take condition and in rounds 6+ of the Private and Public conditions. Error bars represent standard error of the mean. Externalizing costs leads to large increases in punishment levels. These high levels continue in the One-Round Take condition although punishment has no possible effect on future behavior. The fact that the

[17] Players were told that if the interaction ended before the penalty rounds were up, they would not be charged for the extra rounds. In those analyses, we look at the number of rounds that players chose and not the number of rounds that they ended up administering, as these would mechanically decline as the end of the game got closer.

punishments in the One-Round Take condition are positive and higher than in the Private condition shows that cold glow, as a private benefit to punishment, is a major motivating force of decisions to punish.

Columns 1–4 of Table 2 present regression results that confirm the intuitions presented in the figures. We regress the amount of punishment chosen on a dummy that takes a value of zero for Private and one for Public. Standard errors are clustered by participant. Participants in the Private condition chose smaller levels of punishment than those in the Public condition. This result holds when we control for round effects (column 2).

For our robustness check, we pool the data to tease apart the relative importance of public motives (deterrence and incapacitation) and cost structures in choices of punishment. We regress punishment choices on a dummy for costs being public (Public and One-Round Take conditions) versus private and a dummy for public-good (deterrence or incapacitation) motives (Public and Private conditions) versus the One-Round Take condition. The coefficients on these dummies represent the effects of cold-glow versus public-goods motives in punishment decisions. The first dummy is significantly positive: people choose more rounds of exclusion when the costs are public. The second dummy is negative, smaller in magnitude, but not significant, which implies that non-cold-glow motives play a weak role in punishment behavior in our experiment.[18]

Taken together, our regression analyses confirm that cold glow is a good predictor of responses of punishment decisions to cost structure and that indeed aggregate levels of punishment are above those consistent with Beckerian punishers. We note that other motives appear to exist but cannot explain most of the variation in punishment. We now turn to examining the effects of conditions on decisions to take.

3.3.3. Decisions to Take

Figure 3 shows taking decisions by availability of punishment. Although punishment levels are much higher in the Public than in the Private condition, they have no effect on realized levels of taking. However, potential takers react to the possibility of punishment: with no punishment possible, taking levels are very high. Columns 5 and 6 of Table 2 present our regression results. Taking behavior is significantly higher in the Punishment than in the No Punishment condition (column 5), which shows that general deterrence does matter: only 10–20 percent of participants who were able to take[19] chose to do so, even from round 1. However, there is no difference between the Public and Private conditions (column 6).

We find a slight learning effect in the No Punishment condition. Approximately 70 percent of individuals chose to take in the first round, and by the fifth round, 85 percent of participants chose to take. There is no significant difference

[18] Another possible explanation for the difference in behavior between the One-Round Take and Public conditions is that perhaps it is easier to ex post rationalize punishment decisions in the former than in the latter.

[19] That is, players who were not excluded from the game at the time.

Table 2

Experiment 1: Costs and Availability of Sanctions

	Punishment Level					
	Public versus Private			Costs versus Deterrence: Robustness Check	Taking Behavior: No Punishment versus Punishment	
	All Rounds (1)	All Rounds (2)	Rounds 6–20 (3)	(4)	Sanctions (5)	Costs (6)
Public	1.818*	1.809*	2.113**	2.082**		−.0556
	(.754)	(.754)	(.771)	(.768)		(.0728)
Round		−.0406*				
		(.0186)				
No deterrence				−.988		
				(.780)		
Punishment versus no punishment					−.655**	
					(.0371)	
Constant	1.734**	2.166**	1.394**	1.420**	.841**	.219**
	(.455)	(.530)	(.457)	(.458)	(.0256)	(.0625)
N	1,067	1,067	782	902	2,407	1,067

Note. Results are clustered at the subject level. Standard errors are in parentheses.

* $p < .05$.

** $p < .01$.

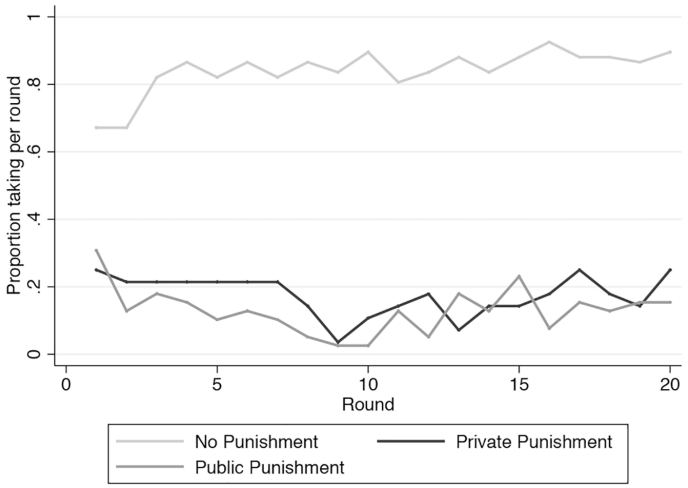

Figure 3. Experiment 1: percentage of participants choosing to take, by condition

between experimental sessions. Thus, although general deterrence did lower taking levels, the extra punishment in the Public condition did not further reduce taking.

3.3.4. Differences in Punishments: Mechanisms

What drives differences in punishment choices across treatments? Table 3 displays the differences across the Private and Public conditions in the number of rounds in which an individual chose not to punish (column 1), and conditional on choosing a positive punishment, the average levels of punishment by subject (column 2). These results are statistically suggestive but not significant at conventional levels ($p < .1$). In the Public condition, individuals are much more likely to opt into using any punishment and, conditional on punishing, give longer punishments. Thus, differences in levels of punishments come from both the extensive and the intensive margins. Table 4 shows the regression of decisions to punish on a dummy that takes a value of one in each round after an individual's choice to punish is implemented. On average, it appears that having paid for punishment does not influence the choice of sentences (column 1). However, the effects are heterogeneous across conditions (columns 2–4): in the Private condition, participants punish significantly less once their punishment has been chosen. We interpret this as a form of sticker shock.

4. Experiment 2: Probability of Apprehension

Our second experiment tests how differences in the probability of apprehension affect punishers' and potential norm breakers' decisions. If punishers and

Table 3
Experiment 1: Individual Differences in Punishing, by Condition

	Rounds with No Punishment (1)	Average Sanction, If > 0 (2)
Public	–3.151+	1.408+
	(1.777)	(.809)
Constant	7.536**	3.374**
	(1.356)	(.634)
N	67	57

Note. Results are clustered at the subject level. Standard errors are in parentheses.
+ $p < .10$.
** $p < .01$.

norm breakers do not symmetrically react to these changes, socially wasteful levels of punishment may result, since probabilities of apprehension enter into optimally deterring punishments. In particular, if norm breakers respond to these probabilities but punishers do not, low probabilities of apprehension can lead to excessively low levels of punishment. In addition, we compare ex ante and ex post punishment decisions.

4.1. Experimental Design

We used a game to test how both sentences and potential norm breaking are affected by expected punishments. The basic design is as follows: players were divided into groups of three to play a one-shot game. They began with a balance of 80 points. Players were randomly assigned one of three roles: assigner, taker, or target. The rules of the game were known to all players before they began the experiment. The game proceeded as follows: The assigner committed to a publicly known level of penalty units (between 0 and 10); each of these units corresponds to a 10-point sanction. Knowing this level of sanction, the taker decided whether or not to take from the target. If the taker chose to take, he gained 20 points and the target lost 30 points. The taker was caught with probability p. If the taker was caught, he was imposed the sanction chosen by the assigner. The assigner was charged 1 point per 5 points of sanction she assigned.

Our treatments vary the probability that the taker will be caught if he takes: in the high-probability treatment, the taker is caught with a probability of 9/10 and in the low-probability treatment with a probability of 1/3.[20] Final payoffs de-

[20] Some studies in psychology investigate the effects of probability of apprehension on punishment decisions. These studies directly ask participants to compare hypothetical punishments in different scenarios when probabilities of apprehension change (Baron and Ritov 2009) or to assess the relative importance of deterrence or moral motives on punishment decisions (Carlsmith, Darley, and Robinson 2002). In these hypothetical contexts, players state that they do not want to change behaviors on the basis of probabilities of apprehension. Our experiment adds to this literature as an incentive-compatible test of whether punishers respond to probability and deterrence motives. In our games, rules are perfectly transparent and deterring punishments are very easy to calculate.

Table 4
Experiment 1: Length of Punishment—Individual Differences

	(1)	(2)	(3)	(4)
Public	1.887*			1.777*
	(.778)			(.750)
Punishment chosen	.579	−1.214+	1.936*	
	(.640)	(.686)	(.936)	
Taken from				−.369
				(.287)
Constant	1.437*	2.358**	2.789**	1.896**
	(.640)	(.733)	(.587)	(.515)
N	1,067	448	619	1,067

Note. Results are clustered at the subject level. Standard errors are in parentheses.
+ $p < .10$.
* $p < .05$.
** $p < .01$.

pended on choices made by all of the players. Finally, the targets made no choice in our game, but we asked them to enter what they thought would be a fair punishment for a taker who chose to take.

We used the online labor market Amazon Mechanical Turk (AMT) to recruit individuals to play the game for a show-up fee of $.30 and an additional payment depending on points earned, using a conversion rate of 2 points per $.01 at the end of the experiment.[21]

We recruited a total of 340 individuals (mean age of 28.8; 63 percent male) to play the game. Each individual played one role in the interaction. To make sure that all participants understood the experiment, they were first given a set of instructions followed by a three-question comprehension quiz (see the online appendix). If they failed to answer any of the quiz questions correctly, they were not allowed to play the game. Dropping noncomprehenders, we were left with 243 individuals (a 71 percent pass rate).

4.2. Experiment 2: Results

4.2.1. Punishers' Behavior

We now consider the behavior of punishers across conditions. The left-hand side of Figure 4 presents assigners' average punishment levels for each of the probability conditions. (Error bars represent the standard error of the mean.) Mean punishment levels are exactly the same in both treatments: the probability of apprehension is not a parameter to which individuals respond when making punishment choices. The mean punishment level is 4.0 units (40 points) in the

[21] Several recent studies undertake to examine the validity of experimental data collected using Amazon Mechanical Turk (AMT) at stakes of about $1. They find that behavior on AMT matches well with standard laboratory results for economics games (Amir, Rand, and Gal 2012; Rand, Greene, and Nowak 2012) and is based on samples that are more representative of the general population (Horton, Rand, and Zeckhauser 2011; Paolacci, Chandler, and Ipeirotis 2010).

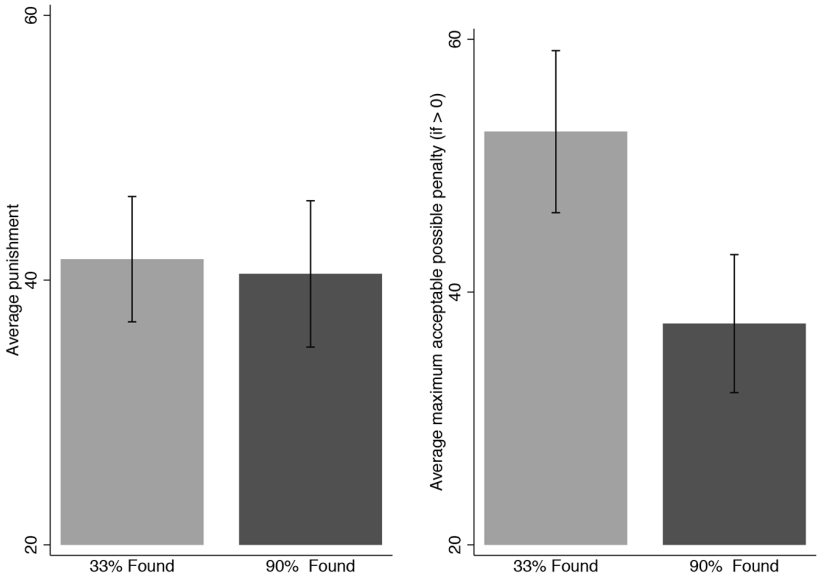

Figure 4. Experiment 2: effect of the probability of apprehension on punishers' (*left*) and takers' (*right*) decisions.

high-probability condition and 4.1 unit (41 points) in the low-probability condition, and the difference is nonsignificant (see columns 1–3 of Table 5). In the Beckerian model of punishment, probability of apprehension is a key feature in determining optimal sentences. Expected punishment is defined as punishment if caught, multiplied by the likelihood that offenders are caught, so changes in probability should be compensated for by changes in punishment. Empirically, experiment 2 shows that this is, in fact, not the case: punishments do not differ across probabilities of apprehension.

4.2.2. Decisions to Take

We find that takers' behaviors, however, do respond to the probability of apprehension on the intensive margin. We use the strategy method to elicit choices to take: takers were asked to enter their maximum acceptable possible penalty (MAPP). This is a number of penalty units such that if the assigner chose a penalty equal to or below this level, the taker would prefer to take. If the assigner chose a larger penalty, the taker prefer not to take. We performed analyses on choices of MAPP to understand takers' behaviors.

A relatively large number of participants (approximately 30 percent) chose a MAPP of 0, which indicates that they did not wish to take under any circumstances in either condition. Column 1 of Table 6 presents our regression results and confirms that there is no significant extensive-margin response. However, focusing on the 70 percent of individuals who entered a MAPP greater than 0,

Table 5
Experiment 2: Punishment Choice, by Probability of Being Caught

	Punisher's Choice			Target's Opinion: Full Sample: Fair Level
	Full Sample		Level If Chose to Punish	
	Punish (1)	Level (2)	(3)	(4)
High	−.131	−.135	.592	.724
	(.0792)	(.738)	(.736)	(.732)
Female	−.0300	−.733	−.703	−.817
	(.0794)	(.739)	(.739)	(.789)
Constant	.935**	4.505**	4.836**	4.600**
	(.0688)	(.641)	(.617)	(.586)
N	81	81	69	80

Note. Standard errors are in parentheses. High = 90 percent chance of being caught .
** $p < .01$.

we find that there is an effect on the intensive margin: as shown in the right-hand side of Figure 4, individuals who chose to take chose different levels of MAPP between probability conditions (mean MAPP for low probability = 5.1 and for high probability = 3.8). Column 3 of Table 6 shows our regression results and confirms that there is a significant intensive-margin response.[22] Unlike punishers, takers respond to the probability of being caught,[23] so the punishment levels chosen are too low to deter many takers in the low-probability condition.

4.3. Control Study: Ex Post Punishments

A key part of our theory is that we allow for both an ex ante (simulating a strategic motive such as deterrence) and an ex post (or "just deserts") component. To assess the size of these components, we ran a control experiment on AMT (n = 194, mean age of 28.9, 63 percent male). The design of the game in our control study is identical, except that the order of moves was switched: takers first chose to take or not, and then assigners chose ex post penalties to assign to takers who were caught. We used the same probability conditions in this study. This has the added benefit of acting as a robustness check on takers' behavior in our original study, in which one possible confound is that takers might have found the strategy method confusing.

Figure 5 and Table 7 present the results.[24] We find that punishers again do not respond to probability of apprehension when choosing levels of ex post punishment (mean punishment for low probability = 3.4 and for high probability = 3.2). Takers, however, do take this probability into account: 25 percent of indi-

[22] We also find a gender effect. Women are less likely to take, and if they are willing to take, they enter lower maximum acceptable punishment levels. This could be explained by higher levels of risk aversion (see Eckel and Grossman 2008).

[23] This also allows us to control for a lack of attention or understanding by participants as the result of a null effect on punishment decisions, as individuals are randomly assigned to roles.

[24] The error bars in Figure 5 represent the standard error of the mean.

Table 6
Experiment 2: Taker's Choice, by Probability of Being Caught

	Full Sample		If Take = 1:
	Take (1)	MAPP (2)	MAPP (3)
High	.114	−.550	−1.700*
	(.0988)	(.721)	(.814)
Female	−.227*	−2.127**	−2.035*
	(.105)	(.767)	(.925)
Constant	.724**	4.116**	5.896**
	(.0785)	(.573)	(.665)
N	82	82	58

Note. Standard errors are in parentheses. MAPP = maximum acceptable possible penalty; High = 90 percent chance of being caught.
* $p < .05$.
** $p < .01$.

viduals take in the high-probability condition, and 43 percent take in the low-probability condition.[25]

We note that in the control condition assigners still choose a positive level of punishment, even though it is a one-time interaction and punishments are privately costly, but probabilities are again not factored in. Levels of punishment are, however, smaller when there is no possibility of deterrence (3.16 ex post versus 4.1 ex ante), these differences being significant only at the 10 percent level. These results are consistent with the differences found in our first experiment between the One-Round Take condition and the Public condition. We conclude that some form of deterrence motives does exist in the punishment choices, but ex post just-deserts thinking seems to be the dominant motivator of punishment behavior in our samples.

4.4. Fairness Judgments

Finally, we look at judgments of fair punishments for takers who are caught from the point of view of the target. Their answers do not appear to differ across conditions (mean fair punishment for low probability, ex ante = 4.3; for high probability, ex ante = 5; for low probability, ex post = 5.3; for high probability, ex post = 5.5).

Column 4 of Table 5 and column 5 of Table 7 present the results of our regression analysis. Unsurprisingly, targets want higher levels of punishment than assigners: this could be driven by differences between second-party and third-party punishment (Fehr and Fischbacher 2004) or because targets do not have to pay for the punishments. Interestingly, neither the order of punishment assignment

[25] This difference is significant, although only at the 10 percent level, because of sample size. The magnitude stays the same—a difference of 20 percentage points—and becomes significant at the 5 percent level when we control for gender.

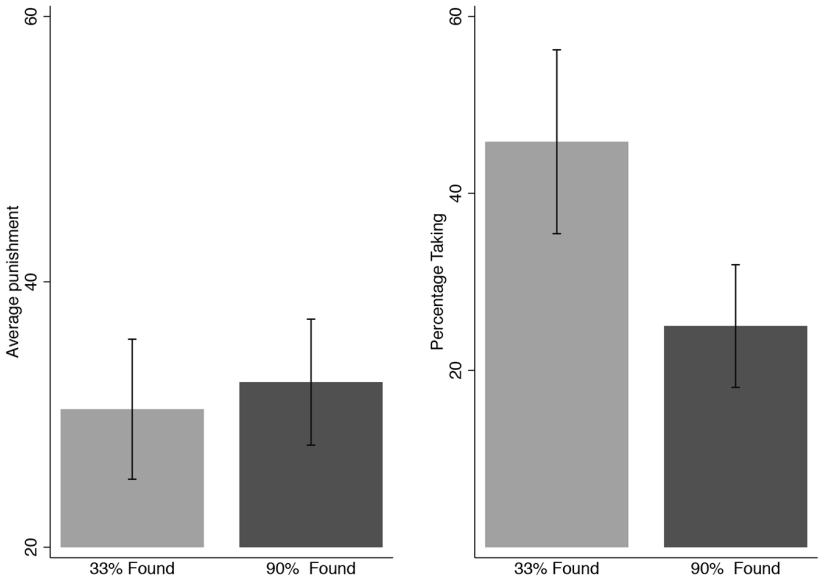

Figure 5. Experiment 2: effect of probability of apprehension on punishers' (*left*) and takers' (*right*) ex post punishment decisions.

nor the probability of being caught changes targets' beliefs about fairness: no extra retribution is demanded when the probability of apprehension is lower. All data taken together, neither punishers nor victims respond to the probability of apprehension when choosing punishment levels, although this parameter seems to matter a lot in the decisions of potential norm breakers.

5. Experiment 3: Crowding Out

Our final experiment asks an individual-level question motivated by our theory: to what extent do the sanction decisions of individuals act as substitutes or complements to their levels of sanction? Our social-level question asks how total levels of sanction change with the introduction of multiple punishers.

5.1. Main Experiment

To answer this question, we ran an experiment on AMT using a sample of 476 individuals (mean age of 29.7; 56 percent male). Participants received a show-up fee of $.50 and an additional payment depending on their earnings during the game, using a conversion rate of 1 point per $.01.[26]

We used a game similar to that in experiment 2 to explore crowding-out be-

[26] Given the average completion time of our experiment and average bonuses, total payoffs amounted to an hourly wage of approximately $8–$10 per hour.

Table 7
Experiment 2: Robustness Check—Punishment and Taking Choices with No Deterrence, by Probability of Being Caught

	Punisher's Choice			Taker's Choice: Full Sample: Take	Target's Opinion: Full Sample: Fair Level
	Full Sample		Level If Chose to Punish		
	Punish (1)	Level (2)	(3)	(4)	(5)
High	.0355	.202	.112	−.251*	.168
	(.0983)	(.741)	(.771)	(.121)	(.850)
Female	.151	−.0533	−.778	−.221+	−.267
	(.0957)	(.722)	(.745)	(.128)	(.867)
Constant	.727**	3.066**	4.189**	.551**	5.456**
	(.0900)	(.679)	(.716)	(.108)	(.771)
N	66	66	54	64	64

Note. Standard errors are in parentheses. Take = player 1 chose to take; High = 90 percent chance of being caught.
** $p < .01$.

havior. Players were randomly assigned to groups of four and started the game with 100 points. Each individual was assigned one role: assigner 1, taker, target, or assigner 2.[27] All rules of the game were known to all players before they began the experiment. Players acted sequentially as follows: assigner 1 committed to a publicly known level of penalty units (0–6); each penalty unit corresponds to a 10-point sanction. Knowing this level of penalty, the taker decided whether or not to take from the target. If the taker chose to take, he gained 30 points and the target lost 40 points. The taker was caught in 3/4 of cases. If the taker was caught, assigner 2 saw the punishment that assigner 1 chose and was given a choice to assign an additional number of penalty units (up to 6). A taker who was caught was imposed the sum of the penalty units chosen by the assigner 1 and assigner 2, and both assigners were charged 1 point per 10 points of sanction they assigned.

Again, although the targets made no choice, we asked them to enter what they thought would be a fair punishment for a taker who chose to take. As in experiment 2, individuals saw the instructions for the experiment and then took a quiz about the rules. Individuals who did not answer quiz questions correctly were not allowed to participate in the experiment. Overall, approximately 70 percent of participants answered the quiz questions correctly, which left us with 73 groups of four players.

Our main variable of interest is the second assigner's choice of level of punishment. As in the previous experiment, we used the strategy method to elicit this preference. Figure 6 presents the average punishment choice of assigner 2 for each possible punishment choice of assigner 1, with error bars omitted. (We calculated statistical significance using clustered regressions because of the correlation of decisions within an individual.) On average, there is no difference across

[27] In the instructions for the experiment, the taker and target are referred to as player 1 and player 2, respectively.

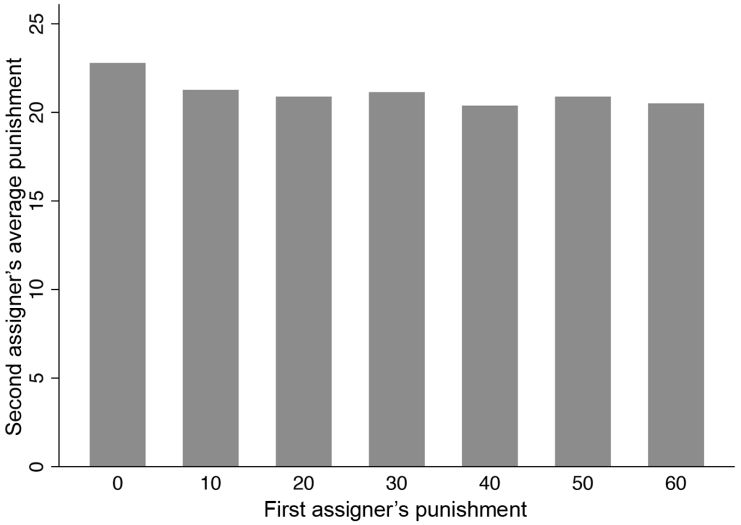

Figure 6. Experiment 3: second assigner's choice of punishment level

the first assigner's choices and thus no evidence of crowding-out behavior on aggregate, as confirmed in the regression analysis (column 1 in Table 8).

We find considerable heterogeneity in individual behavior. Because we use the strategy method, we can look for different behavioral types in our population. Overall, we find that approximately 80 percent of the second assigners can be classified into one of three types: individuals whose sanction choices decrease in the first assigner's choice (partial crowd-out types, 35 percent), individuals whose sanction choices increase in the first assigner's choice (crowd-in types, 25 percent),[28] and individuals whose sanctions do not change as a function of the first assigner's choice (constant types, 20 percent). Individual heterogeneity is not the main focus of this discussion, so we leave it as an avenue for future work. However, we can use this analysis as a robustness check. If we restrict our analysis to the crowd-out types, we still see imperfect crowding out of own punishment by the punishment of another, and we can statistically reject the hypothesis of perfect crowding out even in this restricted subsample (column 2 in Table 8).

We can also look at the average behavior of assigner 1 in this experiment and what the target deems to be a fair punishment. We find that the mean punishment assigned by assigner 1 is 3.02 units (30 points). Combining this with the conditional punishments of assigner 2, we find that the average total punishment given to a player who takes is approximately 5 units, or 50 points. This is 25 percent higher than the mean fair punishment as viewed by targets (mean fair punishment = 42 points).

[28] These individuals may be using the first assigner's decision as a signal of the inappropriateness of taking.

Table 8
Experiment 3: Second Punisher's Choice, by First Punisher's (or Computer) Choice

	Two Punishers		Computer Control: Full Sample: Level
	Full Sample	Crowd-Out Types	
	Level (1)	Level (2)	(3)
Penalty choice of player 1	−.0289	−.569**	
	(.0620)	(.0585)	
High computer penalty			−.355
			(.408)
Constant	2.199**	3.380**	2.053**
	(.237)	(.363)	(.297)
N	553	196	81

Note. Standard errors are in parentheses. High computer penalty = 4.
** $p < .01$.

5.2. *Control Experiment*

Experiment 3 uses a strategy method and a within-subject design to look for the extent of crowd out in punishment. We ran a second study as a robustness check using a between-subjects design without the strategy method. We used AMT to recruit subjects, again dropping those who failed a comprehension quiz. We were left with 243 participants (mean age of 29; 57 percent male) between two conditions.

In our control experiment, players were put into groups of three and assigned a role: taker, target, or assigner. The rules of the game were known to all players before they began the experiment. The game proceeded as follows: The taker decided to take or not from the target. If the taker chose to take, she gained 30 points and the target lost 40 points. The taker was caught in 3/4 of cases. If the taker was caught, she automatically lost c points, where c was varied to be 0 or 40 by condition. If the taker was caught, the assigner could assign up to 6 penalty units, each of which amounted to a 10-point sanction. The assigner was charged 2 points for every 1 penalty unit.

This control experiment allows us to look at crowd-out effects when punishment is assigned by an outside figure instead of another player in the game. Figure 7 and column 3 of Table 8 show the average levels of punishment chosen in the two conditions. (The error bars in Figure 7 represent the standard error of the mean.) Higher levels of punishment assigned by computer do not crowd out individual punishments. Assigners chose slightly lower levels of punishment levels when $c = 40$ than when $c = 0$, but this difference is not statistically significant. It is in any case much smaller than one-for-one crowding out: punishments were on average 2 units in the $c = 0$ condition and 1.7 in the $c = 40$ condition. Thus, total realized sanctions were approximately 20 points when $c = 0$ and 57 points when $c = 40$.

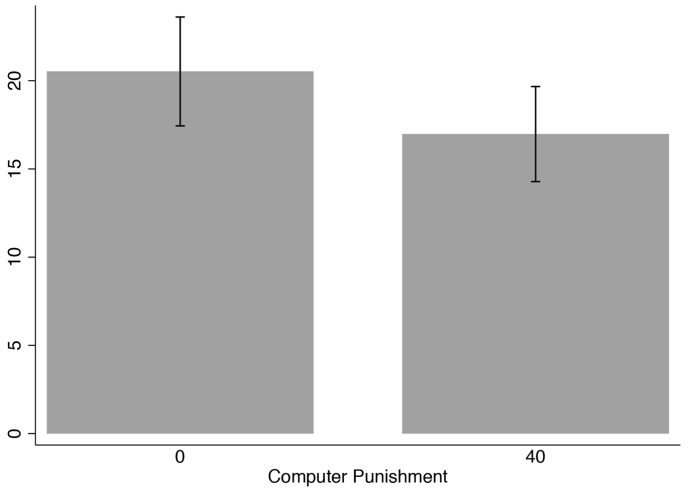

Figure 7. Experiment 3: average levels of punishment when takers who are caught lose no points (*left*) or 40 points (*right*)

We find only a small effect on takers' behavior: 78 percent of takers chose the cooperative action in the $c = 0$ condition, and 85 percent of takers chose cooperative action in the $c = 40$ condition. This difference is not significant, and we attribute the small change to floor effects (recall that takers are caught 75 percent of the time in this control experiment).

This last set of experiments therefore indicates that punishment is not crowded out one-for-one by preset levels of sanctions. On average, there is no effect of preset sanctions on average punishment. We note that there is considerable heterogeneity in this behavior, but we never observe perfect crowding out.

6. Punishment Behavior in the Field

Psychological evidence shows that human punishment behavior is driven largely by blunt, affective motivations. Our lab experiments show that when aggregating these decisions, outcomes may not coincide with Beckerian benchmarks. We now survey some evidence that suggests that cold-glow motivations may have large effects for important outcomes in the criminal justice system.

Demand for punishment for private motives can affect aggregate outcomes through the behavior of elected officials. First, we note that if punishment of offenders is indeed treated by voters as a private good that is provided at public cost, this would lead to demand for punishment even in the absence of clear effects on crime reduction. There is qualitative discussion of this phenomenon: for example, legal sociologist David Garland (2001, p. 133) argues that the most publicized measures (such as three-strike laws or Megan's Law) have little effect on controlling crime but tend to become law because of "their immediate ability to

enact public sentiment, to provide an instant response [or] to function as a retaliatory measure."

In addition to descriptive evidence, causal links have been identified: Berdejo and Yuchtman (2013) analyze changes in the sentencing behavior of judges during election cycles. They find that judicial severity increases when judges are close to reelection and thus under political pressure from constituents and that sentences become lighter immediately afterward.[29] This phenomenon of preelection increase in sentence lengths, immediately followed by a drop, is consistent with a model in which judges' preferences differ from individual voters' decisions, which are driven by the cold-glow heuristic.

Cold glow could also directly affect the behavior of judges and thus outcomes in the criminal justice system. We view that as a less likely source of influence, since judges are specifically trained and make their decisions in a deliberate manner, which perhaps mitigates the effects of cold glow. Recent studies of judicial behavior (Posner 2008; Danziger, Levav, and Avnaim-Pesso 2011) put forth at least some evidence that judges are subject to predictable biases, so it is not impossible that cold glow is a partial motivator of judicial decisions.

In addition, there is evidence in the law and economics literature that argues that individuals may not believe that it is fair to factor the probability of capture into punishment decisions (see Polinsky and Shavell [2000] for a discussion and Sunstein, Schkade, and Kahneman [2000] for two survey-based experiments). Punishers' insensitivity to the probability of capture, an important input in optimal deterrence, is a behavior that cold-glow punishers can display. Further understanding how fair punishment levels are determined is an important direction for basic science as well as for practical considerations.

There has been very little research directly assessing the effect of cost structures on demand for punishment, even though the question of costs of punishment has received attention from policy makers because of the budget crises in many states.[30] Taking our results to the field, Ouss (2014) investigates the 1996 California Juvenile Justice Realignment as a natural experiment. Prior to the realignment, sentences were imposed at the county level, but the costs of incarceration were borne at the state level and were therefore subsidized from the sentencing county's perspective. The realignment resulted in a discontinuous drop in the number of youth committed to juvenile corrections but no discontinuous change in the number of juvenile arrests, which suggests that subsidized corrections had been overused. Similarly, Ater, Givati, and Rigbi (2014) exploit a quasi-experimental change in the costs of arrests in Israel: the responsibility of housing arrestees awaiting trial was transferred from the local police to the prison authority. The authors find a sharp increase in arrests as a result of this policy, consistent

[29] Furthermore, Berdejo and Yuchtman find that this variation is the result of discretionary departure above sentencing guidelines and not greater compliance with them.

[30] In particular, in California, one response has been to house inmates in county jails rather than state prisons, with the argument that this lowers the overall costs of criminal justice.

with an imperfect factoring in of total costs of crime reduction when making decisions to arrest.[31]

Imperfect crowding out of individual punishment preferences by prior punishments could play a role in labor markets. Having a criminal record has an impact on the employability of an individual (Bushway, Stoll, and Weiman 2007; Pager 2007). This can occur through a signaling channel (Rasmusen 1996), where conviction is a signal of a poor worker. However, if cold-glow motives are not crowded out by prior punishments, there may be a second channel for this effect: not hiring an individual who has committed an inappropriate act can function as a sanction. Understanding the relative importance of these channels has important policy implications (for example, policies for shrouding criminal records).

Another interplay between the psychological motivations underlying human punishment behaviors and groups' behaviors can be seen in recent discussions of moblike punishment on social media. Individuals who post socially unacceptable materials on outlets such as Twitter have subsequently been harassed, doxxed (that is, had their personal information such as addresses publicly revealed), and otherwise punished by a massive crowd. This has led in some cases to extreme emotional trauma, job loss, and very high economic, social, and psychological costs for the individuals targeted.[32] While in the context of deterrence motives for punishment this group response seems excessive (surely a smaller response would yield a similar level of deterrence), cold-glow motivations make these events less surprising. Technology makes each individual's act of punishment relatively costless, the public nature of the interactions makes anyone able to join in, and the lack of crowding out of one's desire to punish by others' choices means that no natural mechanism exists to stop the cascade once it begins. Of course, individuals also often receive attention and social status in addition to cold glow for taking part, so a full explanation of why these incidents occur is of independent research interest. However, we argue that the behavioral economics of punishment are front and center in understanding these phenomena.

Individual decisions are a product of many factors: elections involve many nonjudicial dimensions, jurors are prompted to depart from emotions,[33] and exact magnitudes of costs or probabilities of apprehension are generally not known by voters, juries, or judges. In this way, our lab experiments are somewhat artificial. However, they allow us to study, in a controlled environment, punishment choices that are normally hard to observe in the field. We do not argue that experiments are a substitute for traditional empirical analysis but rather a complement; experimental methods form an important part of a larger scientific portfolio. More research is needed, but it seems clear that a richer understanding of

[31] There are many other possible explanations for these results: police officers' effort provision might respond to costs, police evaluations could depend on number of arrests, and so on.

[32] For an illustration of online retaliation, see Ronson (2015).

[33] For example, French jurors verbally pledge that they will "not listen to hatred or malice or fear or affection; [and decide] according to [their] conscience and [their] inner conviction, with the impartiality and rigor appropriate to an honest and free man" (C. pr. pén., art. 304, author's translation).

human psychology can be highly valuable in aiding the understanding of important legal phenomena. We view the growing fields of behavioral and experimental law and economics as important contributors to this understanding.

7. Conclusion

Although many legal scholars and philosophers think of moral reasoning as driven by rational processes, moral psychology suggests that moral behaviors, including punishment of those who break social norms, are driven mostly by emotional reactions, which are then rationalized by conscious processing (Greene and Haidt 2002; Haidt 2001). Using such a blunt psychological mechanism that is motivated by affective factors to make punishment decisions may sometimes collaterally result in social harmony but in other domains can result in either highly inefficient over- or underpunishing. Our lab experiments show little evidence that standard rational motives such as deterrence or incapacitation, which underpin most economics-of-crime models, are major drivers of individual punishment decisions.

We argue that understanding the role that more emotional or automatic mechanisms play in choosing levels of punishments is important to explain outcomes in settings relevant to law and economics, including aggregate outcomes in the criminal justice system. We have presented several possible channels through which cold-glow motivations can affect these aggregate outcomes.

We have primarily compared cold-glow outcomes with material-welfare-maximizing benchmarks, and one may argue that if cold glow is indeed a parameter in a utility function, then cold glow itself could be a legitimate source of individual welfare. However, even if we take cold glow to be a legitimate source of welfare, problems can arise. For example, if individuals can select into the role of punisher, those with the strongest cold-glow motivation might choose to sort into particular positions, and it is unclear that individual maximization will lead to socially optimal outcomes even if cold-glow motivations enter into the calculation of social welfare.

Behavioral and social scientists have increasingly gone beyond studying how aggregate outcomes come about and have begun using their skills to help design rules of the game that achieve normatively desired outcomes (Roth 2002). These types of questions are especially important at the intersections of psychology, law, economics, and institutional design (Hauser et al. 2014; Fudenberg and Peysakhovich 2014): in the case of punishment institutions, effective rules will depend on the psychological motivations of the players. Economics as rule design is a growing and important part of modern social science, and we hope that our results contribute to this important conversation.

References

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal. 2012. Economic Games on the Internet: The Effect of $1 Stakes. *PloS One* 7 (2):e31461.

Anderson, Christopher M., and Louis Putterman. 2006. Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism. *Games and Economic Behavior* 54:1–24.

Andreoni, James. 1990. Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *Economic Journal* 100:464–77.

———. 1993. An Experimental Test of the Public-Goods Crowding-out Hypothesis. *American Economic Review* 83:1317–27.

Andreoni, James, and Laura Gee. 2012. Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision. *Journal of Public Economics* 96:1036–46.

Ater, Itai, Yehonatan Givati, and Oren Rigbi. 2014. Organizational Structure, Police Activity and Crime. *Journal of Public Economics* 115:62–71.

Balafoutas, Loukas, Kristoffel Grechenig, and Nikos Nikiforakis. 2014. Third-Party Punishment and Counter-punishment in One-Shot Interactions. *Economics Letters* 122:308–10.

Baron, Jonathan, and Ilana Ritov. 2009. The Role of Probability of Detection in Judgments of Punishment. *Journal of Legal Analysis* 1:553–90.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76:169–217.

Berdejo, Carlos, and Noam Yuchtman. 2013. Crime, Punishment, and Politics: An Analysis of Political Cycles in Criminal Sentencing. *Review of Economics and Statistics* 95:741–56.

Buckholtz, Joshua W., Christopher L. Asplund, Paul E. Dux, David H. Zald, John C. Gore, Owen D. Jones, et al. 2008. The Neural Correlates of Third-Party Punishment. *Neuron* 60:930–40.

Bushway, Shawn D., Michael A. Stoll, and David Weiman. 2007. *Barriers to Reentry? The Labor Market for Released Prisoners in Post-industrial America*. New York: Russell Sage Foundation.

Carlsmith, Kevin M., John M. Darley, and Paul H. Robinson. 2002. Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology* 83:284–99.

Casari, Marco, and Luigi Luini. 2009. Cooperation under Alternative Punishment Institutions: An Experiment. *Journal of Economic Behavior and Organization* 71:273–82.

———. 2012. Peer Punishment in Teams: Expressive or Instrumental Choice? *Experimental Economics* 15:241–59.

Coffman, Lucas C. 2011. Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics* 3(4):77–106.

Cornes, Richard, and Todd Sandler. 1994. The Comparative Static Properties of the Impure Public Good Model. *Journal of Public Economics* 54:403–21.

Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa. 2009. Accidental Outcomes Guide Punishment in a "Trembling Hand" Game. *PloS One* 4(8):e6699.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences* 108:6889–92.

De Quervain, Dominique J.-F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. The Neural Basis of Altruistic Punishment. *Science,* August 27, pp. 1254–58.

Duersch, Peter, and Julia Müller. 2010. Taking Punishment into Your Own Hands: An Experiment on the Motivation Underlying Punishment. Discussion Paper No. 501. University of Heidelberg, Department of Economics, Heidelberg.

Eckel, Catherine C., and Philip J. Grossman. 2008. Men, Women and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economics Results* 1:1061–73.

Fehr, Ernst, and Urs Fischbacher. 2004. Third-Party Punishment and Social Norms. *Evolution and Human Behavior* 25:63–87.

Fehr, Ernst, and Simon Gachter. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90:980–94.

———. 2002. Altruistic Punishment in Humans. *Nature,* January 10, pp. 137–40.

Fischbacher, Urs. 2007. z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics* 10:171–78.

Fudenberg, Drew, and Parag A. Pathak. 2010. Unobserved Punishment Supports Cooperation. *Journal of Public Economics* 94:78–86.

Fudenberg, Drew, and Alexander Peysakhovich. 2014. Recency, Records, and Recaps: The Effect of Feedback on Behavior in a Simple Decision Problem. Pp. 971–86 in *Proceedings of the 15th ACM Conference on Economics and Computation*. New York: Association for Computing Machinery.

Garland, David. 2001. *The Culture of Control: Crime and Social Order in Contemporary Society*. New York: Oxford University Press.

Greene, Joshua, and Jonathan Haidt. 2002. How (and Where) Does Moral Judgment Work? *Trends in Cognitive Sciences* 6:517–23.

Haidt, Jonathan. 2001. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108:814–34.

Hauser, Oliver P., David G. Rand, Alexander Peysakhovich, and Martin A. Nowak. 2014. Cooperating with the Future. *Nature,* July 10, pp. 220–23.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, et al. 2010. Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science,* March 19, pp. 1480–84.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. Antisocial Punishment across Societies. *Science,* March 7, pp. 1362–67.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics* 14:399–425.

Houser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2008. When Punishment Fails: Research on Sanctions, Intentions and Non-Cooperation. *Games and Economic Behavior* 62:509–32.

Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer, and Ernst Fehr. 2006. Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex. *Science,* November 3, pp. 829–32.

Levitt, Steven D., and Thomas J. Miles. 2007. Empirical Study of Criminal Punishment. Pp. 455–95 in vol. 1 of *Handbook of Law and Economics,* edited by A. Mitchell Polinsky and Steven Shavell. Amsterdam: Elsevier.

McKelvey, Richard D., and Thomas R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10:6–38.

Nikiforakis, Nikos. 2008. Punishment and Counter-punishment in Public Good Games: Can We Really Govern Ourselves? *Journal of Public Economics* 92:91–112.

Nikiforakis, Nikos, and Hans-Theo Normann. 2008. A Comparative Statics Analysis of Punishment in Public-Good Experiments. *Experimental Economics* 11:358–69.

Ostrom, E., J. Walker, and R. Gardner. 1992. Covenants with and without a Sword: Self-Governance Is Possible. *American Political Science Review* 86:404–17.

Ouss, Aurélie. 2014. Incentives Structures and Criminal Justice. Working paper. Univer-

sity of Chicago Crime Lab, Chicago.
Pager, Devah. 2007. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration*. Chicago: University of Chicago Press.
Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5:411–19.
Peysakhovich, Alexander, Martin Nowak, and David G. Rand. 2014. Humans Display a Cooperative Phenotype That Is Domain General and Temporally Stable. *Nature Communications* 5:4939, pp. 1–8.
Peysakhovich, Alexander, and Dave G. Rand. Forthcoming. Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *Management Science*.
Polinsky, A Mitchell, and Steven Shavell. 2000. The Fairness of Sanctions: Some Implications for Optimal Enforcement Policy. *American Law and Economics Review* 2:223–37.
Posner, Richard A. 2008. *How Judges Think*. Cambridge, MA: Harvard University Press.
Rand, David G., Joshua D. Greene, and Martin A. Nowak. 2012. Spontaneous Giving and Calculated Greed. *Nature,* September 19, pp. 427–30.
Rasmusen, Eric. 1996. Stigma and Self-Fulfilling Expectations of Criminality. *Journal of Law and Economics* 39:519–44.
Ronson, Jon. 2015. How One Stupid Tweet Blew up Justine Sacco's Life. *New York Times,* February 12. http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html.
Roth, Alvin E. 2002. The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica* 70:1341–78.
Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science,* June 13, pp. 1755–58.
Shavell, Steven. 1987. A Model of Optimal Incapacitation. *American Economic Review: Papers and Proceedings* 77:107–10.
Singer, Tania, Ben Seymour, John P. O'Doherty, Klaas E. Stephan, Raymond J. Dolan, and Chris D. Frith. 2006. Empathic Neural Responses Are Modulated by the Perceived Fairness of Others. *Nature,* January 26, pp. 466–69.
Sunstein, Cass R., David A. Schkade, and Daniel Kahneman. 2000. Do People Want Optimal Deterrence? *Journal of Legal Studies* 29:237–53.
Sutter, Matthias, Stefan Haigner, and Martin G. Kocher. 2010. Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies* 77:1540–66.
Xiao, Erte, and Daniel Houser. 2005. Emotion Expression in Human Punishment Behavior. *Proceedings of the National Academy of Sciences of the United States of America* 102:7398–7401.
———. 2011. Punish in Public. *Journal of Public Economics* 95:1006–17.