A fine is a more effective financial deterrent when framed retributively and extracted publicly

Tim Kurz*[1], William E. Thomas[1] & Miguel A. Fonseca[2]

[1] *Psychology, University of Exeter*

[2]*University of Exeter Business School*

Word count : 6,934 words

Keywords: Punishment; Morality; Behavior change; Economic; Framing

***Address for contact:**

Dr Tim Kurz
Psychology
Washington Singer Laboratories
University of Exeter
Perry Road
Exeter, EX4 4QG
UNITED KINGDOM
Phone: +44 (0)1392 724657
Email: t.r.kurz@exeter.ac.uk

Abstract

Introducing monetary fines to decrease an undesired behavior can sometimes have the counterintuitive effect of *increasing* the prevalence of the behavior being targeted. Such findings raise important social psychological questions in relation to both the way in which financial penalties are framed and the social contexts in which they are administered. In a field experiment (Study 1), we informed participants who had signed up for an experiment that they would be fined if they arrived late. This fine was presented as either compensatory or retributive in nature and as being administered either privately or publicly. We then observed participants' subsequent arrival time. In accordance with our hypotheses, participants' punctuality was only improved (relative to a no-fine control) in response to retributive rather than compensatory fines and when told that fines would be administered publicly rather than privately. In Study 2 we used a scenario method to demonstrate that the greater efficacy of retributively framed fines can be attributed to their presence being less likely to undermine the perceived immorality of transgression than is the case for compensatory fines. We propose a *material promotion-moral prevention (MPMP) theory* to account for our findings and consider its practical implications for the use of financial disincentives to encourage cooperative behavior through public policy in domains such as climate change.

Author Acknowledgements:

The authors would like to acknowledge and thank Mark Levine, Michelle Ryan, Jessica Salvatore, Kim Peters and Lauren Hall, all of whom provided very helpful feedback on earlier versions of this paper.

A fine is a more effective financial deterrent when framed retributively and extracted publicly

Fines are a part of everyday life. Policy makers routinely employ financial penalties and other deterrents to discourage undesirable actions (such as speeding, littering, or tax evasion) under the assumption that the association of behavior with a punishment will render the targeted behavior less attractive (Becker, 1968; Cooter, 1998). Beginning with Skinner (1945) and Thorndike's (1913) seminal work on operant conditioning, a wealth of experimental evidence has accumulated demonstrating that punishments can successfully decrease undesired behavior (Benabou & Tirole 2006; Eek, Loukopoulos, Fujii, & Gärling, 2002; Fehr & Gächter, 2002; McCusker & Carnevale, 1995; Ostrom, Walker & Gardner, 1992; Van Vugt & De Cremer, 1999; Wit & Wilke, 1990; Yamagishi, 1986). However, as was famously highlighted by Gneezy and Rustichini's (2000) now widely-cited field experiment, introducing fines can sometimes produce counterintuitive effects. In their experiment, they had a group of day care centers in an Israeli city impose a fine on parents every time they picked up their child late. Startlingly, the centers that introduced this fine actually experienced a subsequent *increase* in the number of parents arriving late relative to those who introduced no such fine. Similar findings have also been obtained in other field contexts as well as in the social psychological laboratory (Benabou & Tirole, 2003; Fehr & Falk, 2002; Holmås, Kjerstad, Lurås & Straume, 2010; Mulder, van Dijk, De Cremer & Wilke, 2006).

The theoretical account provided by Gneezy and Rustichini (2000) for this highly counter-intuitive finding was that parents simply viewed the fine as a 'price' paid to perform a desirable and convenient behavior (arriving later) rather than as a punishment for wrong-doing. Along similar lines, but from a more social psychological perspective, Mulder (2008; 2009) has

suggested that psychological and behavioral responses to deterrents may be strongly determined by whether the deterrent in question is perceived as retributive or compensatory in nature. Deterrents are perceived to be compensatory when they are seen as a means by which one compensates ('pays') for the negative consequences of one's transgression (e.g., required overtime pay for childcare staff). On the other hand, they are perceived to be retributive when they are seen as a means by which one is punished *because* one has transgressed a moral norm, such as the norm to avoid inconveniencing others (also see Darley & Pittman, 2003). As Mulder (2009) notes, it is only when a deterrent is interpreted as retributive that it is likely to frame the undesired behavior in terms of shared moral standards. Thus, she suggests, the threat of retributive punishment is more likely to produce the desired effect on behavior, particularly if the material cost of the fine is not especially high. When the deterrent is interpreted in a more compensatory way, however, a social actor is more likely to see fines as simply providing an opportunity to compensate the victims of the social actor's behavioral choice (e.g., the child care workers), thereby actually making transgression seem (paradoxically) more attractive under such a system.

Studies examining the more general effects of situational construal on economic behavior have suggested that the extent to which a decision-making context is framed in moral terms can have a marked impact on the decisions that people make. For example, Lieberman, Samuels and Ross (2004) showed that simply introducing a Prisoner's Dilemma paradigm to participants as 'The Community Game' generated twice as much cooperation than when it was introduced as 'The Wall Street Game'. Although these findings demonstrate the potential impact of different 'mindset' frames on decisions within the context of social traps, they do not directly address the capacity for such effects to be elicited in the context of implementing different forms of

sanctions. The empirical work that comes closest to such an insight is that of Tenbrunsel and Messick (1999), in which business students imagined themselves in a hypothetical scenario whereby they were a company production manager tasked with making a decision that was conceptually akin to a two-player Prisoner's Dilemma. Participants were told that they should make their hypothetical decisions regarding whether to cooperate or defect in light of either a) an inspection/sanction regime with low probability of detecting defection and a relatively inconsequential fine for defection (weak sanction), b) a regime that had high detection probability and a substantial fine (strong sanction) or c) no inspection regime. The findings showed that although cooperation rates were highest under strong sanctions, levels of cooperation under weak sanctions were actually lower than when no sanction was present. Tenbrunsel and Messick (1999) suggest that sanctions led to participants adopting a 'business' frame rather than an 'ethical' decision frame – a claim supported by the recorded post-hoc self-reports of participants, in which those in both sanction conditions were more likely to report having adopted a 'business' frame relative to those participants in the no-sanction condition.

Tenbrunsel and Messick's (1999) findings therefore speak to the possibility of different mindsets being invoked by sanctions of different magnitude (also see Mulder, Verboon & DeCremer, 2009) and/or differing levels of detection probability. The primary implication of Tenbrunsel and Messick's model is that if an authority is going to fine at all, then they must fine big or risk undermining cooperation via the removal of a moral motive. However there are many situations in which the implementation of harsh sanctions is not politically palatable for an authority, especially in response to less serious incursions. As a result, what becomes both theoretically and practically crucial is developing an understanding of whether exactly the *same* type or strength of sanction can lead to greatly different levels of co-operation simply by virtue

of it being framed in ways that invoke different mindsets or interpretations of the social meaning of the sanction. The theoretical distinction drawn by various authors between retributive and compensatory frames in terms of the levels of morality that they convey regarding the target behavior (Gneezy & Rustichini, 2000; Mulder, 2009) offers one such potential theoretical handle. However, as yet, there is no direct empirical evidence to support this model. The current studies aim to provide such evidence.

We also seek in the current work to address a second, previously overlooked, social aspect of Gneezy and Rustichini's (2000) methodology — the public or private administration of fines. Specifically, the notice that was posted on the day care center bulletin boards to communicate the fine's introduction to parents included the following statement: "The fine will be calculated monthly, and it is to be paid together with the regular monthly payment" (Gneezy & Rustichini, 2000, p. 27). The strictly private nature of this transaction thus meant the payment of lateness fines did not involve late-coming parents having to face (or anticipate facing) the individuals they had inconvenienced (i.e., the day care workers).

The legal literature is replete with theoretical and philosophical debate regarding the potential efficacy and ethics of the use of public sanctions in the criminal justice system (for examples, see Flanders, 2006; Kahan, 1996; Massaro, 1997; Shemtob, 2013; Whitman, 1998). However this discussion has been predominantly focused on the more extreme end of the spectrum in terms of both the behaviors in question and the sanctions imposed (i.e. public *shaming* as an alternative to prison for criminal offences). By contrast, our focus here lies in empirically examining the potential for the threat of more mild forms of social pressure or disapproval (what Massaro (1997) might instead define as '*guilt*'-inducing) to increase pro-social behavior. Moreover, as Flanders (2006) points out, one of the potential problems with using fines

as a deterrent is their potential 'fail[ure]…along the expressive dimension', suggesting that fines convey 'the message that crime is merely costly behavior, rather than something that society unequivocally condemns' (p.614). Public sanctions have thus usually been suggested as an alternative, more value-expressive, response to crime that has the additional benefit of also causing less harm to offenders and being less costly than imprisonment. However one might question whether public sanctions need necessarily be thought of only as a more value-expressive *alternative* to financial deterrents when the possibility also exists to give financial deterrents themselves a more public flavor.

There is, as yet, relatively little direct experimental evidence for the efficacy of publicly extracted financial deterrents. Xiao and Houser (2011) have demonstrated in a laboratory context that the public implementation of weak financial disincentives in public goods games are more effective in increasing cooperation than the same disincentives implemented privately. Consequently, it is possible that Gneezy and Rustichini's (2000) field results may have also resulted, in part, from the private nature with which the imposed fine was advertised as being extracted. However direct evidence for such effects in field settings is currently lacking. In the current work we empirically explore whether the potential value-expressive failures of financial deterrents might be overcome, at least in the context of encouraging more pro-social behavior, by making the process of their extraction more public in nature.

The current studies

We investigated across two studies whether the amount of behavioral change produced by materially identical financial deterrents can depend upon how such fines are framed and administered, and also the psychological mechanisms by which such effects might be produced. In our first study we conducted a field experiment that investigated the effects on real, observed

behavior of fines framed as either retribution or compensation that were expected to be administered either publicly or privately. In our second study we employed a scenario methodology to explore whether, as we suggest above, the effects of fine framing on behavior might be driven by the differing levels of morality ascribed to a behavior when fine announcements are worded in either retributive or compensatory ways.

Study 1

To empirically test the effects on behavior of the compensatory versus retributive framing and public versus private extraction of fines, our first study involved a field experiment in which we measured individuals' punctuality in response to the threat of a fine imposed for late arrival at an experiment. We hypothesized that individuals would be more punctual in their arrival when the fine was framed as retribution (rather than compensation) and when participants believed that the deterrent would be administered publicly (rather than privately).

Method

Participants and Design. Participants were 205 undergraduate students at the University of Exeter who were randomly allocated to one of six conditions within a 2 x 2 experimental design in which Frame (compensatory vs. retributive) and Context (public vs. private) of fine were both manipulated between participants. In addition, we included two no-fine control groups[1] (one using retributive wording and one using compensatory wording – as outlined below). Of those who initially signed up to participate, 57 (28%) failed to attend their session, leaving a total sample of 148, 96 of whom were female, with a mean age of 20.27 years (SD = 2.67). The number of no-shows did not differ across compensatory and retributive fine frames, $c^2(1, N = 125) = .01$, $p = .92$, or private and public contexts of fine administration, $c^2(1, N = 125) = 1.30$, $p = .25$. Rates of no-show also did not differ between those in the conditions involving a fine and

those in the no fine control, $c^2(1, N = 204) = .95$, $p = .33$. Participants received £5 (~7.75 USD), on average, for their involvement in the experiment.

Materials and Procedure. An initial recruitment email was sent to a number of undergraduate student research participation lists advertising that participants were sought for an 'economic psychology game' study that would take approximately 30 minutes and inviting those interested to sign up online. Would-be participants were also informed that they would be remunerated with a £2 attendance fee for showing up and would be given the opportunity to earn up to an additional £6 (£3 on average) during the study, dependent upon decisions made in the experimental game by both themselves and the other participants in their session.

All sessions presented to participants started at 9am on weekdays (with day of session being counter-balanced across experimental conditions). Upon signing up for their choice of session, participants received an email confirmation. This email highlighted the importance of arriving in time for the start of the session at 9am and stipulated, for those in the fine conditions, that participants who were more than 15 minutes late would be fined their £2 attendance fee as a result. The reason for this fine was manipulated to be either compensatory or retributive within the text of the email. Those in the compensatory fine condition were informed that "lab space is in high demand in the department at this time of year" and thus it was important that participants arrive on time because "arriving late may hamper our ability to complete the session, which will have financial implications for the research project" and that "as a means to compensate for this, latecomers will forfeit their £2 show up fee if more than 15 mins late". This constructed the fine as partly offsetting the negative implications of participants' wrongdoing. Those in the retributive fine condition were informed that it was important that they arrive on time simply because "latecomers will cause large inconveniences" and that 'for this reason, latecomers will

forfeit their £2 show up fee if more than 15 mins late'. This version of the email thus highlighted why late attendance was 'wrong' - with the fine being justified *solely* in terms of it being reflective of the wrongness of arriving late.

To test for the possibility that differences in the wording of these two frames (*other* than just the ability of the fine to compensate for wrong-doing) might affect participants' perceptions of the seriousness of late coming, we pilot tested the email wordings on a sample of undergraduate students from the same university. These 60 participants (32 female, M_{age} = 21.17, SD_{age} = 3.60) were approached on campus and told that we were looking for students to participate in a short study to help us decide how best to communicate the importance of turning up on time to experiments in our lab. Participants who agreed to participate were provided with a screen shot of either the email sent to participants in the compensatory condition or the email sent to those in the retributive condition, but without the actual possibility of a fine being mentioned in either email. Thus, those in the retributive pilot condition simply read that it was important that they arrive on time because latecomers will cause large inconveniences, whereas those in the compensatory pilot condition simply read that that lab space is in high demand in the department at this time of year and thus it was important that participants arrive on time because arriving late may hamper our ability to complete the session, which will have financial implications for the research project. Removing the actual fine from the email allowed us to pilot test whether basic differences in the wording of the two emails might be responsible for different perceptions of the importance of arriving on time, as opposed to the specific framing of the fine itself as compensatory or retributive. Participants in both conditions were then asked to rate their level of agreement on a 7-point Likert scale with 3 items (α = .72) measuring the level of harm they felt would be caused by coming late to such an experiment (e.g., 'If I arrived late to this

experiment it would result in negative consequences for others'; 'It wouldn't be a major problem if I arrived late to this experiment' [reversed]).

Results of this pilot indicated no significant difference in perceived harm caused between retributive and compensatory wordings (t (58) = 1.56, p = .12), and in fact showed that the harm caused by lateness was, if anything, perceived to be higher in the *compensatory* wording condition (M = 5.66, SD =1.02) than in the retributive wording condition (M = 5.21, SE =*1*.23). Thus, we can be reasonably confident that any superior punctuality by those in the retributive fine condition cannot simply be attributed to higher perceptions of the seriousness of lateness driven by basic differences in the wording of the two emails.

The social context of the fine's administration was also manipulated within this same email. Those in the private fine condition were told that "Late fines will be privately deducted from participants' payments", whereas those in the public condition were told that "Due to the nature of the experiment, late fines will be publicly deducted from participants' payment in front of other players".

For participants in the no fine control condition, half received an email with exactly the same retributive wording outlined above regarding the importance of being on time, except that there was no possibility of a fine mentioned. The other half of the control participants received the same wording as those in the compensatory condition, except that there was again no mention of a possible fine.

Participants in all conditions were requested to reply to the email they received to confirm that they had read and understood the message. All participants complied with this request. In order to maximize the salience of the manipulations prior to arrival at the experiment, a final reminder email was sent to participants 24 hours before their scheduled session, which repeated

this same manipulation of the independent variables. This also had the benefit of reducing potential noise in the data produced by any small and randomly distributed differences across participants in the time lag between the initial confirmatory email and the date of their particular session.

The exact amount of time before or after 9am that each participant arrived at the experimental session was recorded to the nearest 10 seconds. No participants in the fine conditions arrived after 9.15am, removing the need to actually fine any participants their attendance fee. After arrival, participants took part in an economic psychology game (the specific procedure of which is not outlined here because it constitutes a separate study). Upon completion of the game, participants were given their monetary reward, which depended on their performance in the game in addition to their £2 attendance fee. Participants were thanked for their contribution and fully debriefed regarding the study's aims.

Results

Across the sample of participants who showed up to the study, 34% arrived after the requested time of 9am, with a mean arrival time overall of 1.64 minutes before 9am (SD = 4.77). Due to there being no significant differences between the two no-fine control conditions in either number of participants arriving after 9am, $c^2(1, N = 54) = .78$, $p = .38$, or mean arrival time, $F(1,53) = .29$, $p = .64$), these two conditions were collapsed into one single no fine control condition[2].

To assess our hypotheses we first tested if the dichotomous measure of whether participants arrived before or after the requested time was affected by the framing of the message and the social context in which the fine would be (putatively) administered. As shown in Figure 1, and in support of our hypothesis, participants were half as likely to arrive late in the retributive

fine condition (20%) compared to both the compensatory fine condition (40%, Fisher's exact test (1, *N*=93), $p = .02$), and the no fine control (41%, Fisher's exact test (1, *N*=100), $p = .02$). Furthermore, participants were also almost half as likely to arrive late in the public condition (21%) than they were in either the private condition (39%, Fisher's exact test (1, N= 93), $p = .049$), or the no fine control (41%, Fisher's exact test (1, *N*=101, $p = .03$). The cumulative effect of these two independent variables represented the difference between only 9% of participants in the public-retributive condition transgressing and arriving late compared to 48% of participants arriving late in the private-compensatory condition (see Table 1).

We also conducted a two-way ANOVA[3] to test whether framing and context of administration affected participants' exact arrival time, given that arriving *well* before the start time represents a different behavioral response to a participant taking the risk of being late by choosing to arrive just before 9am. Moreover, arriving *substantially* late is clearly quite a different response to arriving only a matter of seconds late. As shown in Figure 1, and in accordance with our initial hypotheses, we observed a significant main effect of frame, $F(1,143) = 3.79$, $p = .05$, $\eta^2 = .03$, with pairwise comparisons indicating that arrival times for those in the retributive condition ($M = 2.95$ minutes before 9am, $SD = 4.24$) were significantly earlier ($p = .04$) than those in the no fine control ($M = .99$ minutes before 9am, $SD = 5.33$) and also marginally significantly earlier ($p = .056$) than those in the compensatory condition ($M = 1.08$ minutes before 9am, $SD = 4.37$). There was no significant difference between compensatory and control conditions ($p = .91$).

Similarly, we observed a significant main effect of context of administration, $F(1,143) = 6.96$, $p = .01$, $\eta^2 = .05$), with pairwise comparisons again revealing that those in the public condition arriving significantly earlier ($M = 3.28$ minutes before 9am, $SD = 3.68$) than those in

both the private condition (M = 0.75 minutes before 9am, SD = 471, p = .01) and those in the no fine condition (M = 0.99 minutes before 9am, SD = 5.33, p = .02). There was no significant difference in arrival time between the private condition and the no fine control (p = .79). There was also no significant interaction between fine frame and context of fine administration on exact arrival time ($F(1,143) = 0.40$, p = .53). However, again, the *cumulative* effect of these two independent variables represented the difference between participants arriving over 4.5 minutes early (on average) in the public-retributive condition compared to arriving (on average) only 6 seconds before the agreed start time in the private-compensatory condition (see Table 2).

Discussion

The results of Study 1 provide empirical evidence to support the theoretical claim made in past literature (e.g., Fehr & Falk 2002; Gneezy & Rustichini, 2000; Mulder, 2009) regarding the potential for financial deterrents to be ineffectual when framed and interpreted in a compensatory, rather than retributive, fashion. The study also provides field evidence for the greater behavioral effect of fines threatened to be extracted publicly, rather than privately. However a key element of the theoretical model that we outlined in our introduction was that retributively framed fines should be more effective *because* they are more likely to lead people to conceive the undesired behavior in terms of shared moral standards. Our second study set out to provide empirical evidence for this proposed psychological process mechanism.

The typical way to demonstrate that a particular psychological process is driving the effect of a manipulation on a behavioral outcome measure is to measure that process as a mediator variable within the behavioral study itself. This was not an option in the context of the current field paradigm, however, due to our behavioral measure (arrival time at the session) having to be performed by participants prior to us having an opportunity to measure anything

else (such as morally wrong lateness was seen to be it was to be). As a result, any extent to which a participant's measured perceived immorality of late arrival might correlate with their own actual arrival time would risk being completely confounded by potential cognitive dissonance effects (Festinger, 1957). In essence, if a participant had just arrived very late to our study and we then asked them how immoral it is to arrive late, cognitive dissonance theory tells us that this participant would be motivated to reduce their state of cognitive dissonance by deciding that it is not so bad after all to be late. Obviously, the opposite would potentially be true for participants who had arrived early. Thus, any observed 'mediation' within such a paradigm would be theoretically muddy at best and completely spurious at worst. As a result, in our second study we developed a slightly less direct way to demonstrate that the differential effects of retributive and compensatory framings of financial deterrents might be driven by their relative abilities to construe the targeted behavior in moral terms.

Study 2

In this study, participants read a scenario about another student research participant, Robin, who had turned up late to a study for which he had been told late attendance would result in a similar fine to those threatened to our real participants in Study 1, with this fine again being framed in either a retributive or compensatory way. We subsequently measured the extent to which participants who read this scenario perceived the target individual's lateness to constitute a moral transgression. We hypothesized that the target's lateness would be seen as less of a moral transgression when performed in response to a fine with a compensatory frame relative to when the fine was retributively framed.

Method

Participants and Design. Participants were 90 undergraduate students at the University of Exeter who were randomly allocated to one of three conditions: a retributive fine, a compensatory fine or a no fine control. Of these 90 participants, half (45) were female. The mean age of participants was 20.67 years old (*SD* = 2.62) and 75% identified their ethnicity as 'White British'.

Materials and Procedure. Participants were approached on campus by a research assistant who asked if they would be willing to complete a very short survey study examining 'how people perceived the conduct of others'. Those who agreed to participate were then asked to read a written scenario about another undergraduate student ('Robin') from the same university. It was explained that Robin had signed up to participate in an experiment in the psychology department that would involve him and a group of other students arriving at the lab at the same time to play an economic decision making task.

It was then explained that, upon signing up for the study, Robin had been sent an email by the researcher running the study ('Jason Bell') providing more details regarding Robin's participation. Below this explanatory text was inserted a screen shot of an email from Jason Bell to Robin, which appeared (visually) just as it would in a standard Outlook email platform. Participants in the retributive condition were presented with exactly the same email that was received by participants in the retributive condition of our field experiment (Study 1) and those in the compensatory condition viewed the same email as that received by those in the compensatory condition of the field experiment. For those in control condition, half received the email from the no fine control in the field experiment that used the retributive wording but with no mention of a fine and half received the equivalent no fine control email from the field experiment that used the compensatory wording but with no mention of a fine. These were

combined into a singular no fine control condition (as was the case in the field experiment). Underneath the screen shot of the email it was explained (in text) that, two days after receiving the email in question, Robin turn up *late* for the group experiment.

Having read this scenario, participants were then asked to answer a series of questions about 'their perception of Robin's conduct'. This involved a 5-item scale (α = .78), which we used to measure the extent to which they perceived Robin's lateness to represent a moral transgression (e.g., "It was morally irresponsible of Robin to arrive late"; "Robin's lateness seemed very socially irresponsible to me"; "I wouldn't judge Robin for being late"[reversed]). Participants indicated their level of agreement with each statement on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree).

In addition, we also presented participants with 3 items (α = .81) measuring the extent to which they thought that late arrival to the experiment in question would have been common (i.e. normative) among all the students who signed up (e.g. "I think that most other students probably also showed up late for Jason's experimental session"). This measure was included to ascertain whether the effect of fine frame in our field experiment might, alternatively, have been attributable to participants having had a perception that late arrival would be more normative when the fine was framed in a compensatory (verses retributive) way, rather than the more specific mechanism of moral construal of the target behavior. Upon completion of the questionnaire participants were fully debriefed as to the aims of the study and were given a small confectionary item to thank them for their time.

Results

We conducted two one-way ANOVAs with fine frame as the independent variable in both cases and perceived moral transgression and perceived norm of lateness as the dependent

variables in respective cases. In both ANOVAs we controlled for the effects of age, gender and ethnicity (White British vs. other) by including them as covariates in the analyses.

Perceptions of moral transgression. We observed a significant main effect of fine frame on participants' perception of Robin having performed a moral transgression, $F(2,83) = 5.67$, $p = .005$, $\eta^2 = .12$, which is depicted in Figure 2. Participants were actually most likely to perceive Robin's lateness as a moral transgression when no fine was attached to late coming ($M = 4.64$, $SD = .89$), a point to which we will return in our discussion. Although perceptions of moral transgression in the retributive condition were slightly lower ($M = 4.28$, $SD = 1.13$) than in the no fine control, pairwise comparisons revealed this difference to be non-significant ($p = .20$). In line with our hypotheses, however, perceptions of moral transgression in the compensatory condition ($M = 3.69$, $SD = 1.28$) were, significantly lower than both the no fine control ($p = .001$), and the retributive condition ($p = .04$).

Perceived norms of lateness. In general, participants indicated that they would not expect lateness to the experiment in question to be particularly normative, with overall mean scores being 2.77 ($SD = 1.24$) on the 7-point scale. These perceptions were almost completely unaffected by the fine frame, with virtually no effect at all of fine frame being observed on perceived norms of lateness, $F(2,83) = 0.02$, $p = .97$, $\eta^2 = .001$.

Discussion

The findings of this scenario study support our theoretical claim that retributively-framed financial deterrents encouraged cooperative behavior more effectively in Study 1 because they frame defection behavior as more of a moral transgression. In line with our theoretical predictions, participants in this study were less likely to perceive the target individual's late arrival to the hypothetical experiment as a moral transgression when a fine was framed in a

compensatory way, than when that same fine was framed retributively, or when no fine was present. Conversely, perceptions of moral transgression in the face of a retributively worded fine did not differ significantly from the no-fine control. Furthermore, these effects of fine frame did not extend to more general perceptions of how likely people in general would be to cooperate in the face of the fine. Rather, the compensatory frame seemed to more *specifically* undermine the extent to which defection was seen as a moral transgression.

General Discussion

The question of when and why the introduction of a financial deterrent might lead to either positive or negative effects on the targeted behavior has been a topic of theoretical debate within both social psychology and behavioral economics. Our field experiment (Study 1) provides the first direct empirical demonstration that materially-identical deterrents can have markedly different effects on real, observed, behavior as a function of whether the deterrent is framed in terms of it compensating for the offenders' wrong-doing or retributively punishing their violation of social or moral standards. Our findings show that participants who were presented with a late fine framed in retributive terms were twice as likely to arrive on time, and arrived significantly earlier than those who were not threatened with any form of fine. However, when exactly the same £2 fine was presented to participants in a way that suggested its payment might serve to compensate for the wrong-doing, this fine led to virtually no impact on participants' behavior whatsoever. In fact, levels of punctuality in response to this fine were indistinguishable from the no fine control. In our second study we provided evidence (via a scenario methodology) that this difference in effect of the compensatory and retributively worded fines was indeed potentially attributable to their differential capacity to construe the behavior in question in moral terms. We showed in Study 2 that when participants were presented with a scenario in which a target

arrived late in the face of a either a compensatory fine, a retributive fine or no fine at all they were less likely to morally judge this lateness when the fine was framed as compensation (relative to both a retributive fine and no fine).

A Material Promotion-Moral Prevention (MPMP) theory of financial deterrents

The one sense in which the results of our scenario study might initially appear to differ slightly from those observed in the field study relates to the positioning of the moral judgment results for the no fine control relative to the two fine frame conditions. To recap, in our field experiment we observed an *enhancement* of punctuality in the retributive condition relative to the no fine control, rather than an *undermining* of punctuality (relative to control) in the compensatory condition. In our scenario study, in contrast, the compensatory condition *did* appear to undermine ascribed morality (relative to control), with the retributive frame being similar to the control in this case. Indeed the scenario condition in which there appeared the most ascription of moral transgression to the target for being late was when lateness came with no potential fine attached to it. Although these observations of retributive frames enhancing co-operation in Study 1 and compensatory frames undermining morality in Study 2 may at first glance seem slightly discordant, upon further theoretical reflection this is perhaps less the case. It actually makes sense that participants might morally judge a target more for being late when this lateness came with no potential punishment attached that might help 'make up for' for transgression. In a sense, Robin was seen as doing something bad and 'getting away with it', hence our participants in the no fine condition subsequently judged him maximally harshly on the moral dimension. What this suggests is that *all* fines may perhaps be perceived to be at least slightly compensatory in terms of how we subsequently morally judge those who transgress (including ourselves), with judgment being dealt out in particularly sharp measure when one (as

the saying goes) 'gets away with murder'. However, in relation to the behavioral decisions of those choosing to cooperate or defect in the face of such fines, there is obviously a counter-veiling, materially driven, effect that motivates people to avoid the material cost of having to pay the fine. In short, fines potentially provide a material *incentive* that promotes cooperation, but a moral *disincentive* (or license to defect) that can prevent the fine from increasing cooperation.

When conceptualizing fines in this way, the results of our two studies are actually rendered highly theoretically concordant. Participants in our field study had a material incentive to plan to be punctual (i.e. not risking the loss of £2). When this fine was framed retributively, the counter-veiling moral disincentive (license) that the presence of such a fine presented was drastically reduced because there was little sense in which paying the fine was construed as making up for this lateness. Thus, the retributive condition led to a *promotion* of punctuality relative to when no fine was used. In the compensatory condition, the same material incentive to be punctual was still present; however the moral licensing of lateness provided by the compensatory fine offset this in a way that *prevented* an increase in co-operation. Thus, these two counter-veiling forces cancelled one another out such that the effect of the compensatory fine was indistinguishable from the no fine control. In the scenario study, however, no material incentives were present for our research participants in any of our conditions. In effect, that study was only tapping into the morality side of our proposed *material promotion-moral prevention (MPMP)* model. It therefore makes sense that, in this case, we would simply see an undermining of the level of moral judgment applied to the act of defection (i.e. lateness) in the compensatory condition relative to a no fine control, with this being far less the case when the fine was framed retributively.

Our proposed MPMP model also offers a potential explanation for the discrepancy in direction of effects relative to the control group for our field experiment in comparison to the classic Gneezy and Rustichini (2000) childcare center field study. In that original study, a (potentially compensatory) fine was shown to actively undermine cooperative behavior relative to no fine. A potential explanation posited by our model is that the material incentive for our (largely unwaged) undergraduate student research participants of ensuring they did not lose their £2 show-up fee may simply have been greater than was the 10-shekel disincentive to avoid lateness provided by Gneezy and Rustichini to the (potentially time poor but materially affluent) parents in their field study. Thus, the moral license provided by Gneezy and Rustichini's fine may simply not have been as highly offset by a counter-veiling material incentive as was the case in our study. This would explain the actual reduction of co-operation when the (potentially compensatory) fine was present in their study as compared to the mere failure of our compensatory fine to produce any increase in co-operation.

In addition to demonstrating the importance of how financial deterrents are framed, our findings also empirically demonstrate, for the first time in a field setting, the greater impacts of financial disincentives implemented in a public rather than private context. We show that only the threat of a publicly administered fine brought about the desired change in behavior of our participants relative to the no fine control condition, further strengthening similar claims recently made on the basis of behavior observed in public goods games conducted in the laboratory (Xiao & Houser, 2011). Thus, it would appear that the threat of a more publicly extracted fine might act as a more powerful incentive for cooperative behavior. Although there exists a range of arguments to be had about whether it is ethically appropriate to incorporate more public sanctions into the judicial system for more serious offences (Flanders, 2006; Kahan, 1996;

Massaro, 1997; Shemtob, 2013; Whitman, 1998), the current findings do at least provide evidence that even making the payment of financial disincentives relating to minor incursions more public can be effective in amplifying behavioral change.

The 48% versus 8% difference in rates of late arrival between the private-compensatory and public-retributive conditions in our field experiment highlights starkly the practical importance for policy makers of considering both a) the extent to which a financial disincentive is likely to be perceived as signaling a moral standard rather than as simply a price that one can pay for the convenience of defection, and (b) the extent to which the context of a financial disincentive's delivery is public or private in nature. A real world policy context that may be informed by these findings is the attempt to tackle global climate change by using financial disincentives to discourage people from engaging in activities involving a large carbon footprint. An important question becomes whether the effectiveness of simply placing a 'price' on carbon may be somewhat limited (or even potentially undermined) by the extent to which consumers may interpret such measures in a more compensatory way (i.e., "I don't have to feel bad about my high energy consumption because I've made up for it by paying the tax"). Moreover, one should also consider the extent to which such price-based systems do not make individuals or organizations in any way publicly accountable for their carbon-producing actions.

This study provides an overdue piece of empirical evidence supporting the theoretical suggestion (e.g., Fehr & Falk 2002; Gneezy & Rustichini, 2000; Mulder, 2009) that behavioral responses to financial disincentives may depend on whether such policies are interpreted as signaling moral standards rather than the opportunity to 'pay for' negative communal effects of one's actions, which we argue supports our proposed *Material Promotion-Moral Prevention (MPMP)* theory of the effect of financial disincentives on cooperation. Moreover, our findings

highlight the benefits of financial disincentives administered in a more public fashion. Both sets of findings present important practical considerations for any policy maker seeking to implement a system of financial disincentives to encourage more communally beneficial behavior among members of a collective.

Notes

[1]The no-fine control groups were collected exactly 12 months after the treatment conditions in the same weeks of the academic year using entirely equivalent mailing lists (with a new first year student cohort). The no-fine and treatment conditions were found to be equivalent on all measured demographics (Gender: fine = 63% female, control = 68% female; Mean Age: fine = 20.26, control = 20.30; Political orientation measured on a 7-point scale from right to left wing: fine = 4.00, control = 3.91; Ethnicity: fine = 68% White British, control = 62% White British).

[2] It should be noted, however, that (as would be predicted on the basis of our pilot data reported in the method) participants were actually slightly more likely to arrive after 9am when the no fine condition contained the retributive wording (46%) than when it contained the compensatory wording (35%) and exact mean arrival time was also slightly later in the retributively-worded, no fine condition (0.67 minutes before 9am) relative to when the compensatory wording was used (1.36 minutes before 9am).

[3] One extreme outlier (> 3 standard deviations earlier than the mean) was removed prior to analysis.

References

Becker, G.S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy,* 76, 169-217.

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies, 70,* 489-520.

Benabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review, 96,* 1652-1678.

Cooter, R. (1998). Expressive law and economics. *Journal of Legal Studies, 27,* 585-608.

Darley, J. & Pittman, T. (2003). The psychology of compensatory and retributive justice. *Personality & Social Psychology Review, 7,* 324-336.

Eek, D., Loukopoulos, P., Fujii, S., & Garling, T. (2002). Spill-over effects of intermittent costs for defection in social dilemmas. *European Journal of Social Psychology, 32,* 801-813.

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. *European Economic Review, 46,* 687-724.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137-140.

Festinger, L. (1957). A Theory of Cognitive Dissonance. California: Stanford University Press.

Flanders, C. (2006). Shame and the meanings of punishment. *Cleveland State Law Review, 54,* 609-635.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies, 29,* 1-17.

Holmås, T. H., Kjerstad, E., Lurås, H., & Straume, O. R. (2010). Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay. *Journal of Economic Behavior & Organization, 75,* 261-267.

Kahan, D. (1996). What do alternative sanctions mean? *University of Chicago Law Review, 63,* 591-653.

Liberman, V., Samuels, S,. & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner dilemma game moves. *Personality and Social Psychology Bulletin, 30,* 1175-1185.

Massaro, T. (1997). The means of shame: implications for legal reform. *Psychology, Public Policy, and Law, 3,* 645-704.

McCusker, C., & Carnevale, P. J. (1995). Framing in resource dilemmas - loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes, 61,* 190-201.

Mulder, L. (2008). The difference between punishments and rewards in fostering moral concerns in social decision making. *Journal of Experimental Social Psychology, 44,* 1436-1443.

Mulder, L. (2009). The two-fold influence of sanctions on moral norms. In D. De Cremer (Ed.), Psychological perspectives on ethical behavior and decision making (pp. 169–180). Charlotte, NC: Information Age Publishing.

Mulder, L., van Dijk, E., De Cremer, D. & Wilke, H. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology, 42, 147-162.*

Mulder, L. B., Verboon, P., & De Cremer, D. (2009). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology*, *39*, 255-269.

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

Shemtob, Z. (2013). Democracy on display: A case for public sanctions. *The Howard Journal of Criminal Justice*. DOI: 10.1111/hojo.12025

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270-277.

Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly, 44*, 684-707.

Thorndike, E. L. (1913). Educational diagnosis. *Science, 37*, 133-142.

van Vugt, M., & De Cremer, D. (1999). Leadership in social dilemmas: The effects of group identification on collective actions to provide public goods. *Journal of Personality and Social Psychology*, *76*, 587-599.

Whitman, J. (1998). What is Wrong with Inflicting Shame Sanctions? *Faculty Scholarship Series*. Paper 655. http://digitalcommons.law.yale.edu/fss_papers/655

Wit, A., & Wilke, H. (1990). The presentation of rewards and punishments in a simulated social dilemma. *Social Behavior, 5*, 231-245.

Xiao, E., & Houser, D. (2011). Punish in public. *Journal of Public Economics, 95*, 1006-1017.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*, 110-116.

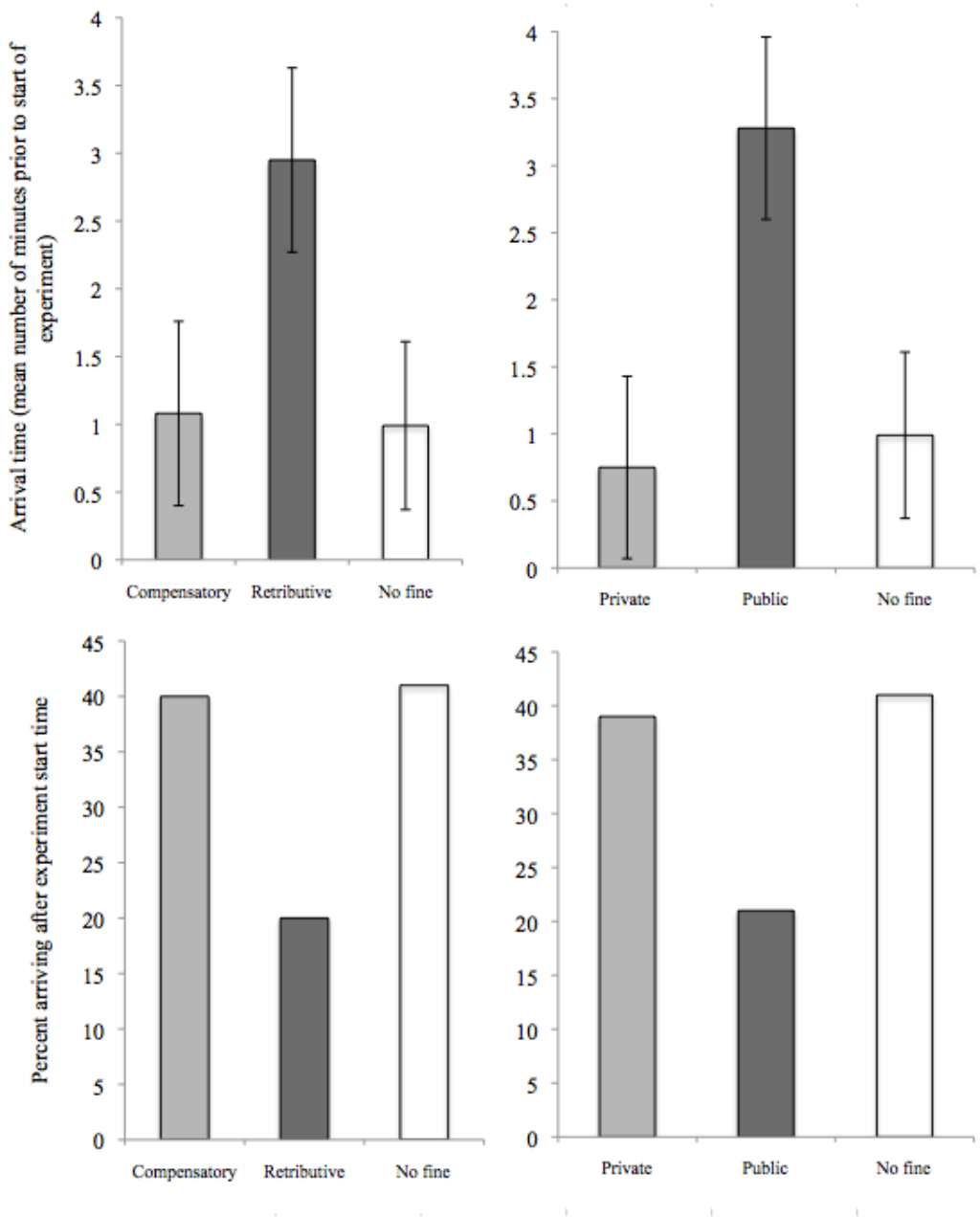

Figure 1. The effect of the fine frame (compensatory vs. retributive) and the setting of the fine (private vs. public) on both mean arrival time and the percentage of participants who arrived late, relative to the no fine control (error bars represent 1 SE above/below mean).

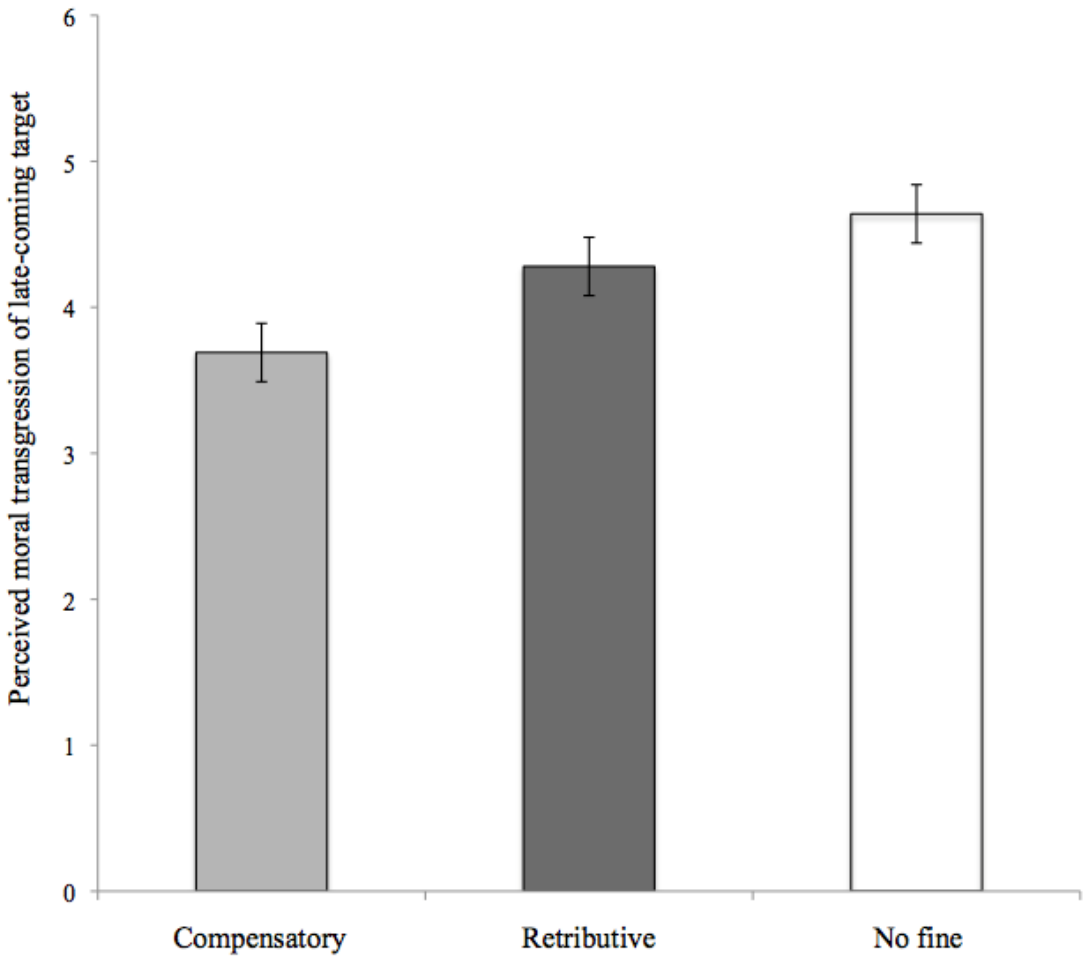

Figure 2. The effect of the fine frame (compensatory vs. retributive) on participants' perception that the target's lateness represented a moral transgression, relative to the no fine control (error bars represent 1 SE above/below mean).

Table 1
Percentage of participants arriving late as a function of fine frame and context of administration.

	Retributive Frame	Compensatory Frame
Public administration	8.7%	33.3%
Private administration	30.4%	47.8%

Table 2
Mean arrival time in number of minutes prior to requested 9am start, as a function of fine frame and context of administration (SDs in parentheses).

	Retributive Frame	Compensatory Frame
Public administration	**4.52** (3.24)	**2.04** (3.74)
Private administration	**1.38** (4.60)	**0.11**(4.83)