A.I. Additional analysis

In the paper, we use the actual behavior in each experiment to estimate a *separate* set of parameters for that experiment. However, a stronger test of the predictive power of elicited norms, when combined with the simple utility function in Equation 1, involves making predictions across experimental populations, i.e., using the parameters obtained from one experiment or set of experiments to predict behavior in another experiment.

Predicting behavior in Lazear et al. (2012), and List (2007) using Experiment 2

We use the elicited norm ratings for the Lazear et al., and List experiments (from Figures 3 and 5 in the paper respectively), the estimated coefficients from model 1 in Table 3 ($\beta = 0.656, \gamma = 1.858$), which were obtained using only data from our Experiment 2, to generate predictions using the same logistic choice model. The predicted distributions are shown in Figures A1a and A1b. In both cases, the changes in behavior when comparing treatments are generally consistent with the observed patterns, respectively, in Figures 2A and 4A that are reported in the paper. In the case of the sorting treatment, Figure A1a predicts that many subjects will choose to opt out and that the frequencies of all amounts shared will decrease dramatically. While the prediction slightly underestimates the frequency of opting out, it is nevertheless highly consistent with the behavior observed in the experiment. In the case of the List experiment, the prediction corresponds less precisely to the data. However, Figure A1b nevertheless captures the main behavioral result in List's experiment, that the introduction of the take $1 alternative generally shifts the distribution of amounts shared downwards, and particularly decreases the frequency of those sharing half of the endowment. Thus, important treatment effects in both experiments are captured by our approach, even when we use parameter estimates obtained from another experiment.

Figure A1a. Predicted distributions of amounts shared in standard vs. sorting treatments using parameter estimates from Experiment 2 (Table 3, Model 1 in the paper)

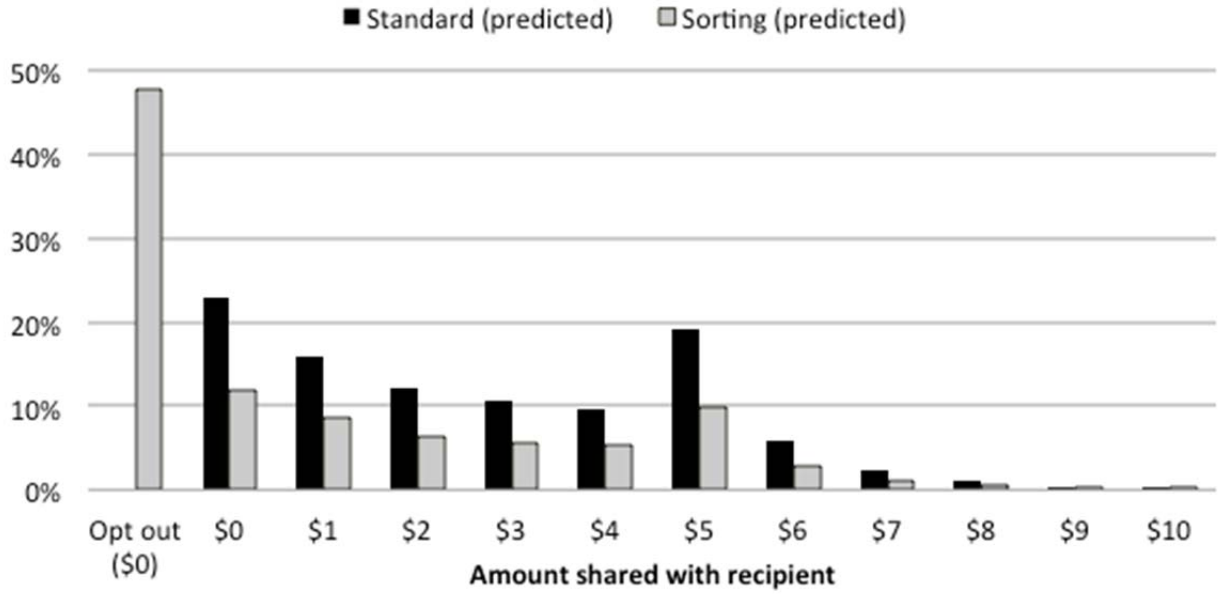
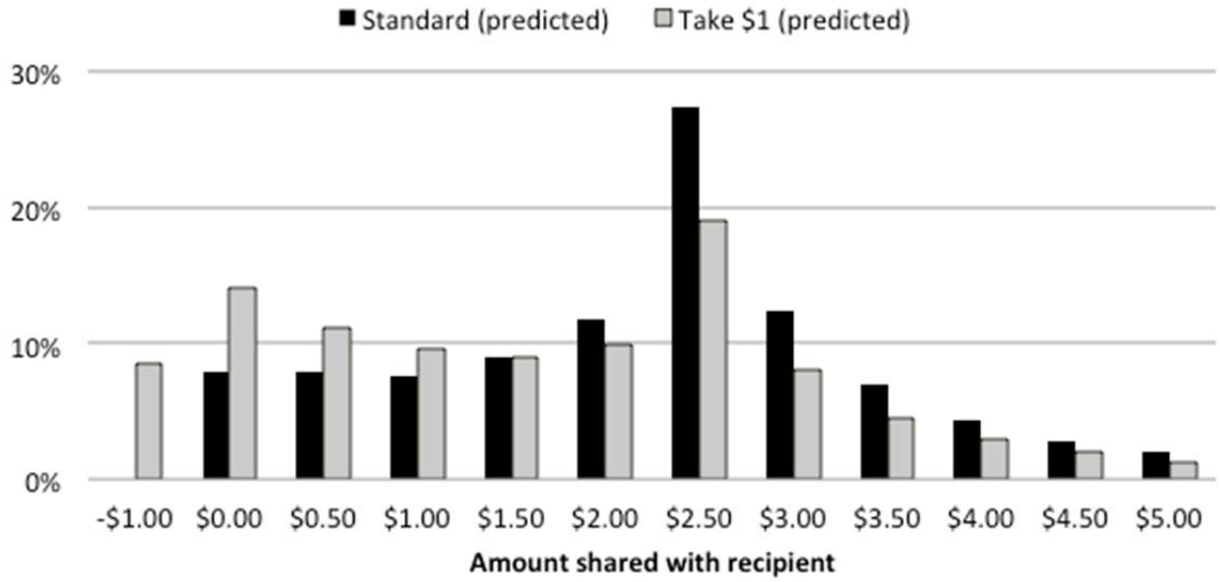


Figure A1b. Predicted distributions of amounts shared in standard vs. take $1 treatments using parameter estimates from Experiment 2 (Table 3, Model 1 in the paper)



Predicting information acquisition in binary dictator games (Dana et al. 2007)

The final two situations for which subjects provided appropriateness ratings in Experiment 1 were two variants of a binary dictator game studied by Dana et al. (2007, henceforth DWK). The DWK experiment found that when dictators could remain ignorant about the consequences of their action for the other person, they often did so in order to behave self-interestedly.[1] More precisely, in a Baseline condition, dictators chose between a ($6,$1) option, labeled "A", and one with payoffs ($5,$5), labeled "B." A majority of subjects in the role of dictator (74 percent) chose the equitable and efficient option B.

In a "Hidden Information" treatment, dictators were told their own payoffs ($6 for choosing A and $5 for choosing B), but not those of the other person. They were told that the two possible payoffs for the recipient were $5 and $1, but not which payoff was associated with which action choice. Therefore, the actual payoffs could be either A:($6,$1) and B:($5,$5) in "scenario 1," as in the baseline condition, or A:($6,$5) and B:($5,$1) in "scenario 2," in which case choosing A maximized the dictator's payoffs, minimized inequality, and also yielded the highest joint payoffs. The actual payoff scenario had been determined by a coin flip prior to the experiment. Subjects in the Hidden Information condition could select to reveal the true payoffs, easily and costlessly, by clicking a button, or could choose to make a choice without finding out the other person's payoffs. In this Hidden Information treatment, significantly fewer subjects chose the fair option B, even though it was always at least as attractive as in the baseline condition (and often more attractive, when the payoffs were flipped). For example, in the case where the underlying payoffs were identical to those in the baseline (scenario 1), only 38 percent of cases resulted in a choice of B ($5,$5), which is significantly lower than the 74 percent in the baseline. Almost half of subjects (44 percent) chose not to acquire the information about the other player's payoffs, even though it was costless to do so, and almost all of these subjects chose A.

To measure the social appropriateness of key actions in this experiment, we presented subjects in Experiment 1 with both the baseline and hidden information variants.[2] For the baseline, they rated the appropriateness of choosing action "A" (yielding payoffs of ($6,$1)) or

[1] This result was replicated by Feiler (2007), Larson and Capra (2009), and Grossman (2009).
[2] The data we use here we collected only in the sessions conducted in Michigan. In the original Pittsburgh sessions, we collected simplified data (not analyzed here) that allowed us to verify that social norms could qualitatively predict the results. See Krupka and Weber (2009) for this earlier analysis using only the Pittsburgh data.

"B" (yielding payoffs of ($5,$5)). The action labels and payoffs were described similarly to the original DWK experiment. The first two rows in Table A1 show the mean ratings of social appropriateness for each of these actions. Not surprisingly, choosing action A in the baseline is socially inappropriate while choosing action B is socially appropriate.

Table A1. Mean ratings of social appropriateness from Experiment 1 for Baseline and Hidden Information Binary Dictator game (Dana, et al., 2007)

Treatment	Action	Monetary payoffs	Mean social appropriateness
Baseline	Choose A (a_A^{Base})	$6, $1	-0.705
	Choose B (a_B^{Base})	$5, $5	0.968
Hidden Information	Don't acquire payoff information – choose A ($a_{no,A}^{HI}$)	$6, $?	0.175
	Don't acquire payoff information – choose B ($a_{no,B}^{HI}$)	$5, $?	0.119
	Acquire payoff information – choose A in scenario 1 ($a_{yes,1,A}^{HI}$)	$6, $1	-0.737
	Acquire payoff information – choose B in scenario 1 ($a_{yes,1,B}^{HI}$)	$5, $5	0.960
	Acquire payoff information – choose A in scenario 2 ($a_{yes,2,A}^{HI}$)	$6, $5	0.793
	Acquire payoff information – choose B in scenario 2 ($a_{yes,2,B}^{HI}$)	$5, $1	-0.765

For the Hidden Information variant, the two possible sets of payoffs were described, as well as the dictator's opportunity to acquire the hidden payoff information. Subjects then rated six possible actions that the dictator might take.[3] These actions, and the associated mean ratings of social appropriateness are presented in Table A1. Note that not acquiring payoff information, and then selecting either choice, is very close to being neither socially inappropriate nor appropriate (mean rating between 0.118 and 0.175). Thus, remaining willfully ignorant is a

[3] Note that we elicited ratings over the final actions, and not over complete strategies available to the dictator. The descriptions of the situation and actions were much easier to explain to subjects in this way. Moreover, this seems appropriate, as social norms are likely to be stronger over the actions one actually takes, as opposed to those one might take as part of a strategy profile.

strategy that can yield high monetary payoffs ($6, if the dictator selects action A), while being more socially appropriate than choosing action A in the baseline (0.175 vs. -0.705). Thus, the elicited norm ratings can explain why dictators who choose to act fairly in the Baseline treatment, by selecting action B, might prefer willful ignorance in the Hidden Information treatment, where they can select action A and obtain the highest personal payoff ($6) by taking an action that is not socially inappropriate.

Conducting the parameter estimation for this experiment that we did for the other three experiments in the paper is not straightforward.[4] We can nevertheless more carefully explore the extent to which our simple framework and the elicited social norms can qualitatively explain behavior, using the parameters we estimated from the other three experiments analyzed in the paper, in model 7 of Table 3 ($\beta = 0.750$ and $\gamma = 1.856$). To do so, we begin by constructing all the possible strategy choices available to the dictator in the two treatments, replacing the actions in our earlier analyses (a_k) with corresponding strategies (s_k), which may include combinations of actions depending on realized uncertainty. For each strategy, we obtain the corresponding (expected) utility, based on Equation 1 in the paper and on the elicited ratings in model 7 of Table 3.

In the Baseline treatment, this is straightforward, as the dictator can only choose actions A or B, and the monetary payoff ($\pi(s_k)$) and social appropriateness ($N(s_k)$) of the strategies are the same as for the actions in the first two rows of Table A1. Thus, for example, choosing the selfish action A in the Baseline treatment yields utility of $u(s_A^{Base}) = 6\beta - 0.705\gamma$

For the Hidden Information treatment, the utility from strategies that forgo acquiring the payoff information is similarly straightforward. For example, choosing not to acquire the payoff information and selecting action A yields a utility of $u\left(s_{no,A}^{HI}\right) = 6\beta + 0.175\gamma$.

However, for strategies in which a subject acquires payoff information, and makes a choice conditional on the realized payoff scenario, we need to consider the uncertainty faced by the dictator at the time of deciding whether to acquire information – regarding the realized

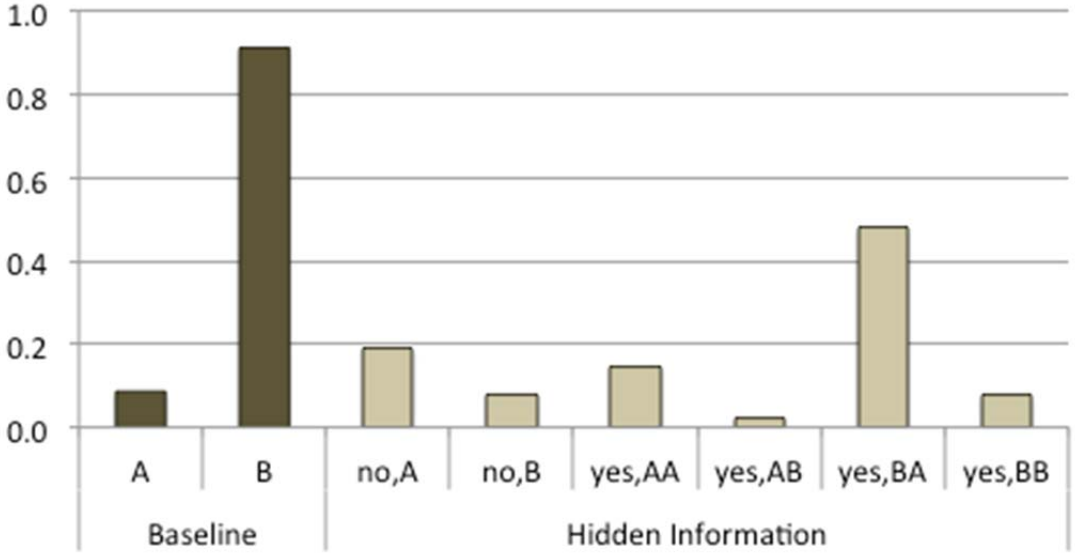
[4] The data from DWK's experiment does not easily lend itself to the kind of estimation in Table 3 for two reasons. First, the experimental data contains only a subject's information acquisition choice and strategy conditional on the realized information in the Hidden Information treatment, which means that it is impossible to know how a subject would have responded to the alternative scenario, conditional on acquiring payoff information. Second, the Baseline treatment contains only a binary choice, which makes identification questionable with two (essentially binary and perfectly correlated) explanatory variables. Moreover, the fact that the dictators face uncertainty requires additional assumptions about risk preferences.

payoff scenario and what action she will ultimately take – in constructing expected utility. Under the assumption of risk neutrality, we construct the expected utility of such strategies based on the values in Table A1. For example, for the strategy, *acquire payoff information, select A if scenario 1 and B if scenario 2* ($s^{HI}_{yes,AB}$), the expected resulting action taken by the dictator is either $a^{HI}_{yes,1,A}$ or $a^{HI}_{yes,2,B}$, with equal probability, based on the realized scenario. Thus, the expected monetary payoff from this strategy is $E\left[\pi\left(s^{HI}_{yes,AB}\right)\right] = \$5.5$ and the expected social appropriateness is $E\left[N\left(s^{HI}_{yes,AB}\right)\right] = -0.751$. We therefore use the expected utility of this strategy $\mathrm{E}\left[u\left(s^{HI}_{yes,AB}\right)\right] = 5.5\beta - 0.751\gamma$, and similarly construct the expected utility of the other three strategies involving the acquisition of payoff information, ($\mathrm{E}\left[u\left(s^{HI}_{yes,AA}\right)\right] = 6\beta + 0.028\gamma$, $\mathrm{E}\left[u\left(s^{HI}_{yes,BA}\right)\right] = 5.5\beta + 0.877\gamma$, and $\mathrm{E}\left[u\left(s^{HI}_{yes,BB}\right)\right] = 5\beta + 0.097\gamma$). We can then use these expected utilities, along with the weights from model 7 in Table 3, to estimate the predicted choice frequencies for each strategy, assuming the same logistic error structure as in our other analysis.

Figure A2 presents the predicted choice probabilities in the two treatments. In the Baseline, the prediction is that the fair option will be chosen with high frequency (91%), which mirrors the actual modal choice in this treatment in the experiment by DWK (74%). Moving to the Hidden Information treatment, the prediction also includes a significant proportion of people choosing to not acquire the payoff information (27%, vs. 44% in DWK's experiment), with the majority of the resulting willfully ignorant choices being A, which is also consistent with the data. Finally, we can compare how the frequencies of B (fair) choices change between the Baseline condition and those cases in the Hidden Information treatment in which the payoffs were identical to those in the Baseline (scenario 1). In DWK, such choices decreased from 74 percent in the Baseline to 38 percent in those cases in the Hidden Information treatment with the same payoffs. Thus, 36 percent of subjects change their behavior between the Baseline and Hidden Information conditions, in the direction of maximizing their own payoff. The comparable decrease in our prediction is from 91 percent in the Baseline to 64 percent under Hidden Information, a decrease of 27 percent.[5] Thus, the proportion of dictators who most starkly change their behavior between the treatments is roughly similar in our prediction and in the data.

[5] The figure of 64 percent is obtained by summing those subjects who do not acquire payoff information and choose B (no,B: 8 percent) and those who acquire the payoff information and choose B under scenario 1 (yes,BA and yes,BB; 48 and 8 percent, respectively).

Figure A2. Predicted choice frequencies in baseline and hidden information treatments of Dana et al. (2007), based on parameter estimates from model 7 in Table 3 in the paper.



Our analysis in this section therefore shows that we can predict, using only our elicited norm ratings from Experiment 1 and the estimated coefficients from other experiments (in Table 3) to predict what is likely to happen in the experiment by DWK. While our prediction is not perfect, and gets some of the specific frequencies wrong, [6] the qualitative patterns of behavior are highly consistent with our predictions. This is in spite of the fact that we are using ("out-of-sample") parameter estimates based only on other experiments. Thus, the analysis here further shows that our approach can be valuable for predicting across experiments.

[6] Overall, the error in our prediction arises mainly because it predicts a higher frequency of fair choices (B), relative to the data, and fewer people are predicted to respond to the treatment. Thus, our prediction overestimates the amount of fair behavior in the baseline and underestimates the behavioral response to the treatment manipulation. This is similar to the error in our out-of-sample prediction of the List (2007) data. Using the parameters estimated only from List's data (model 5 in Table 3), in which subjects appear to care more about money, generates somewhat more accurate predictions for the binary dictator games in DWK.

A.II. Models of social preferences

We consider here the three dictator-game experiments studied in our paper (our Experiment 1 and the experiments by Lazear et al. (2012), and List (2007)), and show that several leading theories of social preferences fail to directly account for observed changes in behavior across the experiments. In particular, we study the preferences present in models of inequality aversion (Fehr and Schmidt 1999), quasi-maximin preferences (Charness and Rabin 2002), guilt aversion (Battigali and Dufwenberg 2007), and social image (Ellingsen and Johannesson 2008). While not exhaustive, these represent a broad set of models that allows us to explore the ability of different kinds of social preferences to explain the changes in behavior across experiments, which we show in our paper are consistent with changing social norms.

We do not study models of reciprocity (Rabin 1993; Falk and Fischbacher 2006; Cox, Friedman and Gjerstad 2007) because a key component of these models – that an individual's social preferences are influenced by the actions of the other player(s) – is absent in dictator games, thus making reciprocity unlikely to apply.[7] We also omit models that are developed primarily to account for one particular experimental finding (Neilson 2009; Dillenberger and Sadowski in press).

Below, we briefly describe each model and, subsequently, the predictions each makes for the experiments that we study. Our main conclusion is that none of the social preference models we review can provide a straightforward explanation for the changes in behavior across all of the experimental treatments we consider in our paper. In particular, models of social preferences based on outcomes alone – as in inequality aversion (Fehr and Schmidt 1999) and quasi-maximin preferences (Charness and Rabin 2002) – do not explain the observed treatment effects.

The most promising of the kinds of social preferences that we consider for explaining the treatment effects is guilt aversion, in which anticipated changes in a recipient's expectations drive changes in the dictator's behavior (as he seeks to avoid feeling guilty because he has fallen short of the expected behaviors). It is not always clear where the anticipated changes in a recipient's exceptions come from – at least in dictator games, where only one player makes a decision. But, if one affords these expectations sufficient flexibility, it is possible to explain the

[7] The model by Cox et al. (2007) includes "status" considerations, in which players' relative status can change social preferences, explaining, for example, the results of Cherry et al. (2002), where relative status changes dictator allocations. This aspect of their model is orthogonal to the dictator experiments considered in our paper, which contain no status manipulation of the kind discussed by Cox et al. (2007).

treatment effects. On the other hand, as we discuss – particularly, in applying guilt aversion to the List (2007) experiment with an additional taking option – it is unclear how or why the expectations should change in a manner that allows this kind of model to explain the observed treatment effects. Therefore, while recognizing that models based on recipients' expectations and dictators' guilt have some explanatory power similar to our interpretation based on social norms, we also note the need to develop improvements for identifying the *source* of such expectations in non-strategic settings such as dictator games. In our paper, we argue that the social norms we elicit – in an incentivized manner – constitute an approach to eliciting expectations about what one should do, and that violating these expectations creates disutility for the dictator.

The Models

For each of the following models, assume that a dictator, *D*, decides how to share an endowment, w_D, by selecting an allocation, $x \in [\underline{x}, \overline{x}]$, for a recipient, *R*. Unless otherwise noted, $\underline{x} = 0$ and $\overline{x} = w_D$, so that the dictator decides how much of the endowment to share. The action choice yields a payoff for the dictator and for the recipient, which are generally $\pi_D = w_D - x$ and $\pi_R = w_R + x$, with w_R, the recipient's initial endowment, usually equal to 0.

In what follows, we consider the preferences of the dictator, based on several models of social preferences that vary in the primary motive driving pro-social behavior.

Inequality aversion (Fehr and Schmidt 1999)

Under inequality aversion (IA), the dictator cares about her own payoff and about the difference between her and the recipient's payoff, according to:

$$U_D^{IA}(x, \alpha, \beta) = \pi_D - \alpha(\max\{\pi_R - \pi_D, 0\}) - \beta(\max\{\pi_D - \pi_R, 0\}),$$

with $\alpha > \beta$ implying that the dictator cares more about disadvantageous inequality than advantageous inequality.

Efficiency and Quasi-maximin preferences (Charness and Rabin 2002)

Charness and Rabin (2002) and Engelmann and Strobel (2004) show that a simple model of preferences for efficiency and for maximizing the lowest payoff to any player (EM) can

explain considerable other-regarding behavior in simple distributional choice tasks similar to the dictator game. More precisely, the preferences are:

$$U_D^{EM}(x, \lambda, \delta\ ) = (1-\lambda)\pi_D + \ \lambda[\ \delta \min[\ \pi_D, \pi_R] + (1-\delta)(\pi_D + \pi_R)],$$

with λ measuring the degree of pro-social orientation and δ measuring the relative weight on maximizing the lowest payoff versus the sum of payoffs.

Guilt aversion (Battigali and Dufwenberg 2004)

If a decision maker experiences simple guilt aversion (GA), then he experiences disutility when disappointing other players by not meeting their expectations. Represented simply, let $E_R[\pi_R]$ correspond to the recipient's belief about how much money she will receive from the dictator. The dictator's preferences are then:

$$U_D^{GA}(x, \gamma, E_R[\pi_R]) = \pi_D - \ \gamma\ max[0, E_R[\pi_R] - \pi_R],$$

where γ measures the dictator's sensitivity to guilt.

Social esteem (Ellingsen and Johannesson 2008)

A decision maker might experience utility from what others think of her, or the social esteem (SE) with which she is regarded. In the case of dictator games, a dictator might share with the recipient because the dictator cares whether the recipient believes the dictator to be concerned with fairness. Specifically, assume that a dictator's type is represented by one of two possible degrees of concern for a recipient's payoff, $\theta_D \in \{\theta_L, \theta_H\}$, with $0 \leq \theta_L < \theta_H < 1$. The dictator then cares about the extent to which she feels esteemed by the recipient, or her "pride," measured by $\hat{\theta}_{RD} = E_{\theta_R}[\sigma(\theta_R)\theta_{RD}|\theta_D]$, where θ_{RD} is the recipient's belief about θ_D after observing the dictator take action x, and $\sigma(\theta_R)$ is the "salience" of the recipient's esteem for the dictator, to capture the property that the dictator cares more about the esteem of someone who is also fair or kind. Assume that $\hat{\theta}_{RD}$ is determined by the recipient's priors on the dictator's type, θ_{RD}^0, and on the action the dictator takes.

The dictator's utility is then:

$$U_D^{SE}(x, \theta_D, \theta_{RD}^0) = \pi_D + \theta_D \pi_R + \hat{\theta}_{RD}$$

The dictator may thus share because she genuinely cares about the recipient, as well as because she wants the recipient to believe that she cares. Importantly, $\hat{\theta}_{RD}$ depends only on priors regarding the different types and on the rational (Bayesian and "reasonable" in the sense of satisfying Cho and Kreps' (2007) Intuitive Criterion) inferences that the recipient can make about θ_D based on the action choice that the recipient observes the dictator take (x) and what this means for payoffs.

The Experiments

We now describe the three dictator-game experiments that constitute the focus of our study, and consider the extent to which each of the above social preference models can explain how behavior changes across treatments. We assume the sensitivities or concern for different components of utility ($\alpha, \beta, \delta, \lambda, \gamma, \theta_D$) are constant. Alternatively, one could too easily explain changes in behavior across treatments by arguing that the intensity of social preferences changes. Moreover, in our main analysis we show that consistent parameters measuring concern for money and social norm compliance can explain the observed treatment differences.

Experiment 2 (Standard vs. Bully)

Our Experiment 2 compares a standard dictator game, where $w_D = \$10$, $w_R = \$0$ and $x \in [\$0, \$10]$ with a bully treatment in which $w_D = \$5$, $w_R = \$5$ and $x \in [-\$5, \$5]$. In both treatments the set of possible payoffs the dictator can implement are identical $\pi_R \in [\$0, \$10]$, and $\pi_D = \$10 - \pi_R$.

Claim: *In comparing behavior between the standard and bully dictator game treatments of Experiment 1, (i) IA, EM and SE do not predict a change in behavior and (ii) GA predicts a change in behavior if the treatment influences $E_j[\pi_j]$, though GA cannot easily account for the complete pattern of data in Experiment 2.*

(i) The dictator's utility in IA and EM incorporates only the possible final payoffs, which are identical across treatments, and the preference parameters, which we assume to be invariant. According to both models, the two treatments are therefore indistinguishable, with a dictator with fixed preferences implementing the same final payoffs in both treatments. For SE, the invariance of θ_D and of the possible π_D and π_R that the dictator can implement imply that $\hat{\theta}_{RD}$ must be equivalent for any equivalent final payoffs across the two conditions. Thus, the dictator faces a choice between identical final payoffs and corresponding degrees of pride across the two treatments.

(ii) In GA, the term $E_R[\pi_R]$ might plausibly be influenced by the treatment. For example, if the recipient's expectations is $E_R[\pi_R] = w_R$, then GA can predict that the dictator will give the recipient a higher payoff in the Bully treatment than in the Standard one. For example, in this case, a dictator with very high γ will leave the recipient with \$5 in the Bully treatment and \$0 in the Standard one. Note, however, that GA cannot account for the increased frequency of \$0 allocations, conditional on not allocating \$5, in the Bully treatment that we observe in the data and for which the elicited norms can account.

Lazear et al.'s Experiment (Standard vs. Sorting)

Lazear et al.'s, experiment compares a standard dictator game, as above, with a sorting variant that includes an additional option, $\emptyset$, to opt out of the game, i.e., $x \in [\$0, \$10] \cup \emptyset$. Choosing this additional option implements payoffs $\pi_D = \$10$ and $\pi_D = \$0$ and leaves the recipient uninformed about the presence of the dictator or the possibility of an allocation.

Claim: *In comparing behavior between the standard and sorting dictator game treatments of Lazear et al. (i) IA and EM do not predict a change in the resulting final payoffs, (ii) GA directly predicts a change in behavior consistent with the observed treatment effect, and (iii) SE does not directly predict the treatment effects but could be modified to do so.*

(i) Under IA and EM, the dictator is solely concerned with final payoffs (π_D and π_R). Therefore, the introduction, in the sorting treatment, of another option that reproduces the $\pi_D = \$10$ and $\pi_R = \$0$ payoffs already present in the standard treatment might lead some dictators to choose

this option, but only those who would choose the same payoffs in the standard treatment. This should therefore not impact the distribution of final payoffs.

(ii) For GA, the sorting treatment introduces an option that eliminates any expectation on the part of the receiver, thus making $E_R[\pi_R] = 0$ and creating an opportunity whereby guilt-averse dictators can obtain $\pi_D = \$10$ without experiencing disutility from guilt. Thus, GA can account for why a dictator who shares positive amounts in the standard variant would prefer to opt out in the sorting treatment.

(iii) For SE, opting out in the sorting treatment would imply that $\hat{\theta}_{RD} = 0$, or is undefined, as the recipient is unaware in this case that the dictator exists and the dictator therefore cannot experience any utility from pride (one could think of this as a situation in which $\sigma(\theta_R) = 0$). Since, by construction, the lowest value that $\hat{\theta}_{RD}$ can take is zero, this makes a dictator indifferent between sharing \$0 in the dictator game and opting out. Note that SE could be modified to accommodate the treatment difference by having a default value of pride greater than zero, which might then allow the dictator to obtain this degree of pride by opting out.[8]

List's Experiment (Standard vs. Take-\$1)

The experiment by List uses a slightly different dictator game as the baseline, in which dictators choose $x \in [\$0, \$5]$. To this standard game, he introduces the additional option of taking \$1 from the recipient in order to create a "Take \$1" treatment, in which $x \in [\$0, \$5] \cup \{-\$1\}$. An experiment by Bardsley (2008) studies a similar treatment. The striking result from List's experiment is that the introduction of this taking option dramatically increases the proportion of subjects opting to share nothing or to take from the recipient, by decreasing the proportion of subjects who share strictly positive amounts when the additional option is not present. As Bardsley (2008) shows, this is inconsistent with giving in dictator games resulting from social preferences based on convex indifference curves over own and the other's wealth.

In testing the ability of social preference models to explain the change in behavior, we focus on a subject's choice to share a positive amount in the standard dictator game – e.g., as

[8] Indeed, Ellingsen and Johannesson motivate their model partly by discussing an experiment very similar to Lazear et al. (by Dana et al. 2006), but do not explicitly show how their model can account for this data.

when $x = \$2.5$ – but nothing in the modified (Take \$1) variant. We refer to this pattern of behavior as the "treatment effect" in List's experiment.

Claim: *In comparing behavior between the standard and take \$1 dictator game treatments of List, and considering the "treatment effect" as the tendency to share \$2.50 in the standard variant and \$0 in the take \$1 variant, IA, EM, GA and SE do not directly predict the treatment effect.*

For IA to generate predictions in the interior of the action space available to a dictator – i.e., sharing half the wealth as when $x = \$2.5$ – requires the introduction of some non-linearity into the utility function. Fehr and Schmidt (1999) propose a modification whereby the dictator's utility is concave in the amount of advantageous inequality, to generate (unique) maxima in the interior of the action space, and they show that the properties of the model are relatively stable to such a modification. Under such a modified utility function, $U_D^{IA'}(x, \alpha, \beta)$, one can predict the treatment effect only if $U_D^{IA'}(\$2.5, \alpha, \beta) > U_D^{IA'}(\$0, \alpha, \beta)$ in the standard treatment, but $U_D^{IA'}(\$0, \alpha, \beta) > U_D^{IA'}(\$2.5, \alpha, \beta)$ in the take \$1 treatment. Since the utility from these outcomes produced by these alternatives is unchanged by the introduction of the additional $x = -\$1$ option, the above two inequalities cannot simultaneously hold across treatments without otherwise changing preferences. A similar argument applies to EM.

For SE, a change in behavior consistent with the treatment effect would require that the dictator feel different degrees of pride, $\hat{\theta}_{RD}$, for taking action $x = \$0$ in the standard and take \$1 treatments, based on a change in the recipient's beliefs about θ_D. This is plausible, but cannot completely account for the treatment effect of subjects changing their behavior from $x = \$2.5$ in the standard treatment to $x = \$0$ in the take \$1 treatment. For example, suppose that a "selfish" type has $\theta_L = 0$, meaning she does not care about the recipient's payoffs. In this case, a dictator who takes the least generous action (sharing \$0 in the standard variant and taking \$1 in the take \$1 variant) might correctly be interpreted as having this type by the recipient. If we assume that dictators who take any action that is less selfish are assumed to have θ_H by the recipient, then a dictator who cares about esteem and has higher concern for the recipient, θ_H, should select the next least selfish action, thus yielding choices by such "generous" dictators of $x > \$0$ in the

standard treatment and of $x = \$0$ in the take \$1 treatment, which looks somewhat like the treatment effect. However, note that if such dictators are motivated mainly by a desire to not be perceived as the bad type, then they should share very little in the standard treatment (an amount close to \$0). The aspect of the treatment effect that SE seems unable to explain, therefore, is the dramatic decrease in generosity by those dictators who share $x = \$2.5$ in the standard treatment.

For GA to explain the treatment effect, the recipient's expectation would have to change dramatically. For example, a dictator sharing $x = \$2.5$ in the standard variant and $x = \$0$ in the take \$1 treatment could be motivated by the receiver's expectations, $E_R[\pi_R] = \$2.5$ and $E_R[\pi_R] = \$0$, respectively in the two treatments. However, this change in expectations does not directly follow from the representation of the two treatments in the model. Imposing structure on how the beliefs are generated from the structure of the game seems unlikely to provide a complete account of the treatment effect. For example, if one assumes that the recipient always expects the mean of the uniform distribution over all the possible amounts she might receive or the midpoint between the best and worst payoffs she could obtain – in both cases, $E_R[\pi_R] = \$2.0$ in the Take \$1 treatment – then the resulting change in behavior is unlikely to account for the large treatment effect. Instead, the recipient's expectations might be based on "focal" amounts, such as \$2.50 in the standard treatment and \$0.00 in the take \$1 treatment, but measuring these expectations begins to overlap significantly with our approach. Therefore, while GA may potentially explain the treatment effect, it does not do so in a straightforward manner.

A.III. Robustness of parameter estimation

Table A2 replicates the analysis in Table 3 of the paper, with one difference. The ratings of social appropriateness here are constructed from the *median*, rather than mean, responses provided by subjects in Experiment 1. Otherwise, all of the analysis is identical. The results in the table demonstrate that our findings are generally very similar under this alternative method for constructing $N(a_k)$.

References

Battigali, P. and M. Dufwenberg. 2007. "Guilt in Games." *The American Economic Review*, 97(2): 170-176.

Charness, G. and M. Rabin. 2002. "Understanding Social Preferences with Simple Tests." *The Quarterly Journal of Economics,* 117 (3): 817-869.

Cox, J., D. Friedman and S. Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior*, 59(1): 17-45.

Dillenberger, P. and P. Sawdowski. in press. "Ashamed to be Selfish." *Theoretical Economics*.

Ellingsen, T. and M. Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *The American Economic Review*, 98(3): 990-1008.

Falk, A. and U. Fischbacher. 2008. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2): 293-315.

Feiler, L. 2007. "Behavioral Biases in Information Acquisition." *Doctoral Thesis, California Institute of Technology, Division of Humanities and Social Sciences*.

Fehr, E. and K. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, 114(3): 817-68.

Grossman, Z.2009. "Self-Signaling Versus Social-Signaling in Giving." *Unpublished manuscript*.

Larson, T. and C. M. Capra. 2009. " Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A comment." *Judgment and Decision Making*, 4(6): 467–474.

Lazear, E., U. Malmendier and R. Weber. in press. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied.*

List, J. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy*, 115(3): 482-94.

Neilson, W. S. 2009. "A Theory of Kindness, Reluctance, and Shame for Social Preferences." *Games and Economic Behavior*, 66(1): 394-403.

Rabin, M. 1993. "Incorporating Fairness into Game Theory and Economics." *The American Economic Review*, 83(5): 1281-1302

Table A2. Conditional (fixed-effects) logit estimation of choice determinants across experiments (includes <u>median</u> appropriateness ratings from Experiment 1 as an explanatory variable)

Behavioral data (experimental treatment)	Experiment 2 (Standard vs. Bully)		Lazear, et al. (in press) (Standard vs. Sorting)		List (2007) (Standard vs. Take $1)		Data from all three experiments	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Monetary Payoff (β)	0.642*** (0.117)	0.631*** (0.123)	0.899*** (0.096)	0.901*** (0.096)	1.505*** (0.357)	1.406*** (0.389)	0.754*** (0.065)	0.715*** (0.087)
Appropriateness rating (γ)	1.585*** (0.317)	1.314*** (0.379)	2.092*** (0.286)	2.065*** (0.300)	1.617*** (0.639)	1.577** (0.616)	1.512*** (0.182)	1.579*** (0.240)
Appropriateness rating X non-standard treatment		0.420 (0.269)		0.098 (0.257)		-0.250 (0.483)		
Monetary payoff X Lazear, et al., experiment								0.129 (0.136)
Appropriateness rating X Lazear, et al., experiment								0.425 (0.422)
Monetary payoff X List experiment								0.568 (0.408)
Appropriateness rating X List experiment								-0.497 (0.892)
2γ/β	4.94*** (0.44)	4.16*** (0.75)	4.65*** (0.29)	4.58*** (0.37)	2.15*** (0.53)	2.24*** (0.59)	4.01*** (0.26)	4.42*** (0.30)
Log-likelihood	-206.3	-205.3	-310.3	-310.2	-125.5	-125.3	-677.6	-653.9
Obs. (subjects)	1166 (106)	1166 (106)	2105 (183)	2015 (183)	816 (70)	816 (70)	4,087 (359)	4,087 (359)

* - p < 0.1, ** - p < 0.05, *** - p < 0. 01; all two-tailed

Bootstrapped standard errors in parentheses. All variables are constructed exactly as in Table 3, except that "Appropriateness rating" is the median rating provided for a particular action.