Measuring Socially Appropriate Social Preferences*

Jeffrey Carpenter† Andrea Robbett‡

July 2022

Abstract

This paper extends the literature on structural estimation of social preferences to account for the desire to adhere to social norms and hide one's true intentions via moral wiggle room. We conduct an experiment to test whether accounting for normatively appropriate behavior allows us to distinguish between preference types who care about outcomes versus adhering to social norms and whether the introduction of moral wiggle room undermines the stability of social preference estimates. We find that social preference estimates are remarkably robust to the inclusion of moral wiggle room. However, the representative agent is strongly motivated by norms and failing to account for this motive in our model causes us to overestimate how much agents care about helping those who are worse off. Using finite mixture models to endogenously identify latent preference types, we replicate previous work finding that the majority of subjects can be classified as strong or moderate altruists when normative concerns are not considered. Accounting for the normative appropriateness of decisions when categorizing participants, however, reveals different motives across types: strong altruists are only marginally concerned with norms while the moderate altruists are highly sensitive to them and, once norms are taken into account, don't care at all about the outcomes of others. Our results thus recast the prior findings in a new light. Rather than the two most common types being strong altruists and moderate altruists, we find that they are better described as strong altruists and norm followers.

JEL Classification: C91, D01, D91, D63, D30, C49
Keywords: Experiment, Social Norms, Social Preferences, Moral Wiggle Room, Structural Estimation, Finite Mixture Models

---

*We are grateful to Guillaume Frechette, Sevgi Yuksel, and participants at the 8th Biennial Social Dilemmas conference at MIT for helpful comments and discussions. The experiment reported in this paper was approved by the Middlebury College IRB and informed consent was obtained. Funding provided by Middlebury College.

†Department of Economics, Middlebury College. Middlebury, VT 05753. jcarpent@middlebury.edu
‡Department of Economics, Middlebury College. Middlebury, VT 05753. arobbett@middlebury.edu

1 Introduction

Across a wide range of contexts, people regularly act in ways that conflict with their narrow material self-interest, even in simple choices where there is little room for confusion. For example, experimental participants in dictator games usually share at least some proportion of the surplus with an anonymous recipient, typically in the range of $20-30\%$ (Engel, 2011). For the last quarter century, the most common means of modeling this behavior is to assume that agents are *other-regarding* and have social preferences that cause them to internalize the outcomes of others (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002; Charness and Rabin, 2002). Indeed, structural estimates of social preferences indicate that the representative participant is willing to sacrifice to help those who earn less (e.g., Andreoni and Miller, 2002; Bellemare et al., 2011; Charness and Rabin, 2002; Fisman et al., 2007; Chen and Li, 2009; Bruhin et al., 2019).[1]

At the same time, experiments indicate that people care, not only about distributional outcomes, but the social appropriateness of their own actions, and whenever there is ambiguity in how actions map to outcomes, some people will exploit this *moral wiggle room* to behave less generously (Dana et al., 2007; Krupka and Weber, 2013; Grossman and Van Der Weele, 2017). Our approach is thus to estimate a model of social preferences that also includes the desire to adhere to social norms and use finite mixture models to assess whether different types are differentially motivated by outcomes or norm adherence. We additionally assess the robustness of social preference estimates to moral wiggle room.

Estimates of social preferences in two-player games most commonly use some variation of the utility specification given by equation (1). This model is due to Charness and Rabin (2002), was extended to include positive reciprocity by Bellemare et al. (2011) and Bruhin et al. (2019), and nests the Fehr-Schmidt inequality aversion model. Player i's utility is a simple weighted average of her own material payoff, x_i, and the other player's payoff, x_j, where the weights may

[1] A recent meta-analysis analyzing 26 papers finds that, on average, participants are willing to give up 40 cents to increase the payoff of someone who earns less by $1 (Nunnari and Pozzi, 2022).

vary by situation.

$$
\begin{aligned}
u_i(x_i, x_j; \alpha, \beta, \delta, \gamma,) = &(\alpha \mathbb{1}_{x_i<x_j} + \beta \mathbb{1}_{x_i>x_j} + \delta \mathbb{1}_{Unkind} + \gamma \mathbb{1}_{Kind}) x_j \\
&+ (1 - \alpha \mathbb{1}_{x_i<x_j} - \beta \mathbb{1}_{x_i>x_j} - \delta \mathbb{1}_{Unkind} - \gamma \mathbb{1}_{Kind}) x_i
\end{aligned} \tag{1}
$$

When player i is behind ($x_i < x_j$), then i puts weight α on the other player's payoff. For $\alpha > 0$, the decision maker is altruistic and willing to sacrifice to help the other player even though that person is already better off. Alternatively, $\alpha < 0$ indicates that i is *behindness averse* – that is, averse to disadvantageous inequality and willing to sacrifice to drag the other player's payoff down closer to her own. In contrast, when player i is ahead, she may weight j's payoff differently, which is captured by the parameter β, where positive β indicates that the decision maker is *aheadness averse* and willing to sacrifice to help someone who is worse off. When $\alpha < 0 < \beta$ the decision maker is both behindness averse and aheadness averse and a slight rearrangement of equation (1) yields the Fehr-Schmidt inequality aversion model (Fehr and Schmidt, 1999).[2] Finally, how generous or competitive the decision maker feels toward someone may depend on how that person has previously treated them and so these weights can further shift by δ or γ if player j has been unkind or kind, respectively.

Over the past 20 years, researchers have structurally estimated this and other similar social preference models either for individual subjects (e.g., Andreoni and Miller, 2002; Bellemare et al., 2008, 2011; Fisman et al., 2007, 2015), for a finite number of predefined preference types (Iriberri and Rey-Biel, 2011, 2013; Conte and Moffatt, 2014; Conte and Levati, 2014; Bardsley and Moffatt, 2007), or by endogenously identifying latent preference types without making any prior assumptions about those types' parameters (Breitmoser, 2013; Bruhin et al., 2019; Van Leeuwen and Alger, 2019). Bruhin et al. (2019) (henceforth BFS) uses finite mixture models to simultaneously estimate the preference parameters in equation (1) for a finite number of types and classify subjects into those types. This method provides a framework for our study. Their estimates are based on the choices of 160 students in Zurich who make 39 binary allocation decisions as dictators and as the second-movers in a games where the first-movers

[2]The Fehr-Schmidt inequality model additionally specifies that $|\alpha| > \beta$ such that the people are more averse to disadvantageous inequality than advantageous inequality. Also note that in the Fehr-Schmidt specification the α value reflects the psychological cost of disadvantageous inequality rather than the weight placed on the other player's payoff, so that the sign is flipped and behindness aversion is captured by a positive α parameter.

have previously acted kind or unkind. Three months later, the same participants returned to the lab and made these same 117 choices again. BFS identify three temporally-stable preference types: about 40% of their subject pool are classified as strong altruists who put large positive weight on the payoffs of others both when ahead and behind (i.e., $\alpha, \beta > 0$) and reciprocate kind (and, to a lesser extent, unkind) acts; 50% are moderate altruists who put lower, but still positive, weight on the payoffs of others and are somewhat negatively reciprocal; and about 10% are behindness averse and put negative weight on the payoffs of those doing better (i.e., $\alpha < 0$) but otherwise don't care about those who are behind ($\beta \approx 0$) or reciprocity.

However, there is also an abundance of evidence that preferences over outcomes alone are not the only motive driving behavior in dictator games. A variety of experiments have found that when the connection between a player's action and the ultimate distributional outcome is obscured, some people take advantage of this moral wiggle room to act selfishly while still maintaining an image as a fair person. Specifically, experiments have found that distribution choices become less generous when dictators can avoid learning the consequences of their decision for the recipient (Dana et al., 2007; Grossman and Van Der Weele, 2017), they have plausible deniability because there is some possibility that they lost their agency and the computer was actually responsible for selecting the selfish outcome (Dana et al., 2007; Andreoni and Bernheim, 2009), they can opt out of the dictator game and just take the entire surplus for themselves (Dana et al., 2006; Lazear et al., 2012), or the option to take away money is added to the dictator's action set such that not giving is no longer the most selfish choice on the menu (List, 2007; Bardsley, 2008).

These changes in behavior across contexts cannot be rationalized if dictators are purely motivated by outcomes, as in equation (1). Instead, Krupka and Weber (2013) proposed that many of these patterns can be explained by a model in which dictators are motivated by the social appropriateness of their actions, in additional to financial outcomes, and introduced the now-ubiquitous method of eliciting social norms using coordination games. Specifically, participants read about a game or decision and rated the social appropriateness of each possible action on a four-point scale ranging from very socially inappropriate to very socially appropriate. The participants knew that they would receive a financial bonus if their rating of a random-chosen action matched the modal rating in their session, such that the ratings reflect, not just personal opinions, but a common perception of the social norms. Using this approach, the authors demonstrated that many of the shifts in dictator generosity could be explained by assuming that people's preferences are represented by a utility function that is increasing in the social appropriateness of their actions.

Our paper additionally relates to work examining heterogeneity in norm adherence. Kimbrough and Vostroknutov (2016) develop a rule-following task to measure norm sensitivity outside the context of social preferences. They find significant heterogeneity in participants' desire to follow rules and that groups of subjects comprised of rule-followers are better able to sustain cooperation. In the context of honesty, Bicchieri et al. (2020) find that about 20% of participants distort their own beliefs about the prevalence of lying when they know that they will later complete a task in which they earn money from being dishonest, suggesting that they dislike violating norms in addition to any intrinsic distaste for lying.

We expand the existing work on the structural estimation of social preferences to include the desire to adhere to social norms and design an experiment to address three research questions. First, do we overestimate outcome-based social preference motives if we don't account for the normative appropriateness of various actions? Second, can we identify different types who are motivated by outcomes versus adhering to norms? To address these two main questions, we conduct a series of 45 dictator games, similar to those used by BFS, adding a parallel treatment in which a different set of participants rate the social appropriateness of each decision using the Krupka-Weber elicitation method. We then incorporate this as an additional variable in our structural estimates to potentially explain behavior.[3] Third, we conduct a robustness check by asking whether including moral wiggle room in our dictator choices affects the social preference estimates and categorization of player types. Here, we include a treatment in which subjects have plausible deniability for implementing the selfish allocation, in the form of a random cutoff rule (à la Dana et al., 2007) that implements the self-interested allocation in the event that the subject does not make a choice within a particular timeframe.

We report several results. First, in an environment without reciprocity or wiggle room,

[3]Because it is not the focus of our study and the estimated weight was always modest in BFS, we exclude the games used to measure a reciprocity motive.

we replicate the three type classifications and preference parameters uncovered by BFS. Second, we find that these results are strikingly robust to the introduction of moral wiggle room. Having established the robustness of these estimates, our primary results concern how our estimates and classification of types vary when we account for social norms. We find that the representative agent is highly sensitive to norms (and, after controlling for distribution outcomes, willing to sacrifice about \$0.80 to \$1 to take an action rated as "very socially appropriate" instead of one rated as "very socially inappropriate.") Once we account for the normative appropriateness of the possible actions, our estimate of β declines significantly, indicating that some of the behavior that would typically be attributed to outcome-based social preferences is actually motivated by the desire to adhere to norms. Without considering norms, we estimate that the representative agent is willing to sacrifice about \$0.52 to increase the payoff of someone worse off by \$1 and this declines to about \$0.32 once norms are taken into account. In other words, norms matter, but they aren't all that matters.

Turning to the finite mixture model estimates, accounting for norms does not influence the number of types that best fit the data or the distribution of subjects into types. However, it does reveal more nuanced and distinct motives for the three types. Without accounting for norms, strong altruists comprise a little more than half of our population, moderate altruists make up a bit less than half the population, and there are some behindness averse participants. While the distribution of types does not change noticeably when we account for norms, the characterization of the most prominent two types does. When we consider norms in our estimates, we find that a little more than half the participants are not just strong altruists – they are strong altruists who also don't care that much about the norm. Further, the inclusion of norms indicates that the moderate altruists are actually better characterized as efficiency-minded norm-followers. These participants adhere closely to norms and don't otherwise care about helping their partner when ahead, but are more efficiency-minded than pure norm following would predict and are still willing to help their partner even when behind.

Finally, we also examine the robustness of our social preference estimates by comparing the main results to those from our moral wiggle room treatment and by considering an alternative interpretation of our results, one that places even more emphasis on norms. The most consistent wiggle room driven difference in our mixture model is in the distribution of types. Overall, when moral wiggle room is introduced, participants are recast as less altruistic. If participants can "wiggle," there are fewer strong altruists and efficiency-minded norm-followers and more of our participants are categorized as behindness averse. Although the effects are not large, so our estimates are mostly stable, this retyping of participants is consistent with the standard literature on moral wiggle room.

In another exercise designed to test the stability and robustness of our estimates, we consider an alternative hypothesis in which all participants are primarily norm-driven but have different subjective perceptions of what the appropriate norm is. In this case, our estimates of α and β would no longer be the parameters of a social preference utility function. Instead they would simply account for participant misperceptions of the norm. To assess the validity of this alternative, we simulate choice data for the participants who just rated the normative appropriateness of the options faced by our dictators using the simple rule that they choose the most appropriate option. The simulated choices result in estimates that are substantially different from our main results and, therefore, indicate that our participants are, indeed, driven by both outcome-based social preferences and a desire to follow social norms.

2 Design

2.1 Allocation decisions

Experimental participants considered the 45 allocation decisions depicted in Figure 1. The circles indicate allocations for the decision maker (horizontal axis) and the recipient (vertical axis) and each line denotes a binary choice between the two allocations it connects. Note that downward sloping lines depict choices in which the decision maker can sacrifice to help the other player, while upward sloping lines indicate that their material interests are aligned but the decision maker can sacrifice to harm the other person. Each pinwheel shape depicts the same 15 tradeoffs, but at different magnitudes: the decision maker is always behind in the games at the top left, always ahead in the games in the bottom right, and her choice determines whether she is ahead or behind in the games in the middle. Among these 45 decisions, 39 are those used by BFS. We added two choices to each pinwheel to capture situations where we believed financial motives and norm adherence were most likely to conflict: when it costs only a little to help the other player a lot.[4]

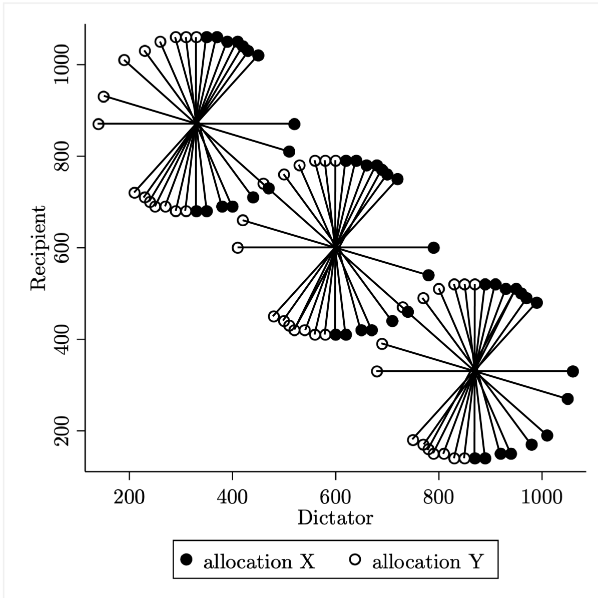

Figure 1: *The experimentally induced allocation choices.*

2.2 Experimental design and procedures

Overall, 598 subjects participated in the experiment, which was conducted on Prolific in April and May of 2022.[5] Each participant either made these 45 allocation decisions *or* rated their social appropriateness and we additionally varied whether the choice permitted moral wiggle room. The four cells to which the participants were assigned are shown in Table 1.

Participants who made allocation decisions faced each of the 45 choices in random order on separate screens. For each choice, the participant viewed the two allocations, labeled X

[4]Participants in our experiment considered only the dictator game version of these choices and not the reciprocity games. We chose to focus on the dictator games to maximize subjects' attention on these 45 choices and because reciprocity preferences did not strongly drive the type classification in the BFS experiment. In particular, they find that "the preference types differ primarily in their distributional parameters" (p. 1056) and when we estimated the finite mixture model on their data without the reciprocity games, the proportion assigned to each type was essentially unchanged (as shown Table A3).

[5]The median age in our sample is 34. 59% of our participants identified as Female. 82% reported that they are White, 6.5% reported that they are Black or African American, and 6.5% reported that they are Asian. 7.7% reported that they are Hispanic. There is considerable variation in reported income: the median participant reported household income between $50,000 and $75,000, while nearly 20% reported income above $100,000 and approximately 13% reported an income below $20,000.

Table 1: Number of participants per treatment cell

	Made Allocation Decisions	Rated Allocation Decisions
No Wiggle Room	194	99
Wiggle Room	205	100
Total:	399	199

and Y, and made a selection. They were paid a participation fee of $1.34 plus the outcome of one randomly selected choice at the exchange rate of 100 points per 1 USD.[6] Those in the No Wiggle Room condition could spend as much time as they wanted on each choice. In contrast, the Wiggle Room condition was based on the *plausible deniability* protocol (Dana et al., 2007; Van der Weele et al., 2014). Each decision was associated with two separate screens. First, the participant saw the choice that they were about to face (the two allocations X and Y) and they had as long as they wanted to consider it. When they were ready, they pressed the continue button. At that point, they knew that the computer could cut them off at a randomly chosen time within 10 seconds. If they had not made their choice by then, the allocation providing the higher monetary payoff for the decision maker (which was always allocation X) would be implemented by the computer. Following the prior experiments, the decision maker was never cut off in under 5 seconds and so in practice they had plenty of time to consider their options and make their choice. However, this treatment gives the participant the option to deliberately choose the self-interested allocation X and still have plausible deniability or to allow the clock to run out and let the computer choose it for them.

A different set of subjects rated the social appropriateness of each of these decisions following the procedures established by Krupka and Weber (2013). For each of the 45 choices, they viewed allocations X and Y and rated each on a four-point scale labeled "very socially inappropriate," "somewhat socially inappropriate," "somewhat socially appropriate," and "very socially appropriate." The raters were paid a participation payment of $3 and additionally earned a bonus of $5 if their own rating on a randomly chosen question matched the modal

[6]The exchange rate is essentially the same as BFS, which was 100 points per Swiss Franc. Participants knew that one of the 45 scenarios would be randomly chosen for payout and it would be randomly determined whether they and their partner were paid based on their own decision or based on their partner's decision. Payments were made as bonuses on Prolific within 48 hours of participation. The average bonus was over $6 and it took a little under 10 minutes to complete the experiment.

rating among those in the same treatment. While rating each choice, those in the No Wiggle Room condition were reminded that the decision maker deliberately chose between X and Y. Those in the Wiggle Room condition were informed of the random cutoff rule, reminded on the rating screen that X could be implemented unintentionally, and instructed to keep this in mind when making their ratings. Following Krupka and Weber (2013), we scale the responses such that the four ratings are scored as -1, $-1/3$, $1/3$, and 1, respectively, and use the empirical mean rating in our estimation, as described in the next section.

2.3 Empirical strategy

We consider two different utility specifications. The first (*Without Norms*) takes player i's utility from allocation X to be a function of the material outcomes for both players, $X = (x_i, x_j)$, as given by Equation (2). The specification is equivalent to that of Charness and Rabin (2002) and BFS when there is no room for reciprocity, and it depends only on player i's behindness aversion parameter α_i and aheadness aversion parameter β_i:

$$u_i(X; \alpha, \beta, \gamma) = (\alpha \mathbb{1}_{x_i<x_j} + \beta \mathbb{1}_{x_i>x_j})x_j + (1 - \alpha \mathbb{1}_{x_i<x_j} - \beta \mathbb{1}_{x_i>x_j})x_i \tag{2}$$

where $\mathbb{1}_{x_i<x_j}$ and $\mathbb{1}_{x_i>x_j}$ are indicators reflecting whether player i earns strictly less or strictly more, respectively, than player j under allocation X.

The second specification (*With Norms*) allows participants to also be motivated by norm adherence, as described in Equation 3. As in Krupka and Weber (2013) and Kimbrough and Vostroknutov (2016), we allow the player's utility from choosing X to increase in the social appropriateness of X, which is denoted by $N(X)$ and given by the mean elicited social appropriateness rating. This specification includes an additional preference parameter, γ, which captures the player's norm sensitivity.

$$u_i(X; \alpha, \beta, \gamma) = (\alpha \mathbb{1}_{x_i<x_j} + \beta \mathbb{1}_{x_i>x_j})x_j + (1 - \alpha \mathbb{1}_{x_i<x_j} - \beta \mathbb{1}_{x_i>x_j})x_i + \gamma N(X) \tag{3}$$

We structurally estimate players' preference parameters using a standard random utility model (McFadden, 1981). Each player i acts as if her true utility from allocations X and Y are given by the expressions in Equations 2 or 3 plus a random error term: $u_i(X) + \epsilon_X$ and $u_i(Y) + \epsilon_Y$. Assuming that the error terms are independent draws from a Type-1 extreme value (Gumbel) distribution with scale parameter $1/\sigma$, the probability that the player chooses X instead of Y is given by:

$$Pr(u_i(X) - u_i(Y) \geq \epsilon_Y - \epsilon_X)$$
$$= \frac{exp(\sigma u_i(X))}{exp(\sigma u_i(X)) + exp(\sigma u_i(Y))}$$

The parameter σ reflects the player's choice sensitivity. When σ is zero the player chooses X and Y with equal likelihood regardless of the underlying utility, and as σ grows larger, the probability of choosing the allocation with higher utility approaches 1.

In what follows, we will first assume that there is a single preference type and estimate the preference parameters, α, β, and γ, and the choice sensitivity parameter σ for this representative agent using maximum likelihood estimation. Specifically, the probability density function is given by:

$$\begin{aligned} f(\alpha_i, \beta_i, \gamma_i, \sigma_i; X_i, Y, \mathbb{1}_X) = \prod_{g=1}^{G} & \frac{exp(\sigma u_i(X))}{exp(\sigma u_i(X)) + exp(\sigma u_i(Y))}^{\mathbb{1}_{X_{ig}}} \\ & \times \frac{exp(\sigma u_i(Y))}{exp(\sigma u_i(X)) + exp(\sigma u_i(Y))}^{1-\mathbb{1}_{X_{ig}}} \end{aligned} \tag{4}$$

where $\mathbb{1}_{X_{ig}}$ is an indicator for whether X is chosen by player i facing decision g in the data. When estimating individual-specific parameters, we have $G = 45$, since for each individual we take the product of the likelihoods (of the data producing that outcome) for each of the 45 data points. When estimating the parameters for the representative agent, we treat all observations as if they were generated by the same agent and we have $G = 45 \times 194 = 8730$ for No Wiggle and $G = 45 \times 205 = 9925$ for Wiggle.

We will then assume that there is a limited number of types, K, and use finite mixture models to estimate the vector of preference and choice sensitivity parameters for each type that maximize the likelihood of observing the decisions in our data. In this case, individual i's contribution to the probability density function is given by sum of the probability densities for each of the K types weighted by the share, π_k, of each type k in the population:

$$\sum_{k=1}^{K} \pi_k f(\alpha_k, \beta_k, \gamma_k, \sigma_k; X, Y, \mathbb{1}_{X_i})$$

Finally, after estimating the fitted values for each type (α_1, ..., α_K, β_1, ..., β_k, γ_1, ..., γ_K, σ_1, ..., σ_K) and the shares of each type ($\pi_1, \ldots, \pi_k$), we can estimate the ex-post probability that each player i is a member of type k using Bayes' rule:

$\tau_{ik} = \frac{\pi_k(f(\alpha_k,\beta_k,\gamma_k,\sigma_k;X,Y,\mathbb{1}_{X_i}))}{\sum_{l=1}^{K} \pi_l(f(\alpha_l,\beta_l,\gamma_l,\sigma_l;X,Y,\mathbb{1}_{X_i}))}$

Following BFS, we estimate the model for $K = 2$, 3, and 4 and select the version with the lowest normalized entropy criterion (NEC). We additionally consider the integrated completed likelihood (ICL). Both are based on a measure of entropy, which is defined as: $EN(\tau) = -\sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{ik} ln \tau_{ik}$. In other words, the entropy is closer to zero when each of the player's ex-post likelihoods assign them to one type with a high likelihood. The NEC normalizes entropy by the difference in log likelihood under the model with K types and the model with a single representative type ($NEC = \frac{EN(\tau)}{L(1)-L(K)}$). The ICL instead adds to entropy the number of preference parameters times lnN and subtracts two times the log likelihood ($ICL = EN(\tau)+$ (number of estimated parameters)$\times lnN - 2L(K)$).

3 Results

3.1 Elicited norms

We first provide an overview of the elicited norms. Across all decisions, the average ratings of X and Y were close to zero (0.0039 and -0.0176, respectively). To understand the relative normative pull across the two options the decision maker could select in any specific dictator game, we consider the average difference in the appropriate ratings of X and Y for each choice, such that positive numbers indicate that X is viewed as more appropriate. These differences in ratings are reported in Table 2 by features of the decision. When the decision maker will end up ahead regardless of which option he chooses (on the left of Table 2), we see that norms have the most bite. Allocation Y is viewed as far more socially appropriate than X when Y is more efficient, reduces inequality, or helps out the other player. Put differently, our raters identify a relatively strong norm to be generous (and not competitive) when one is ahead.

In contrast, when the decision maker is behind (on the right of Table 2), the difference in ratings between X and Y shrinks, but choosing the materially self-interested X is now always more socially appropriate, especially when it would reduce the inequality. In the upper right portion of the table we see that efficiency concerns wane normatively when the dictator is behind. Of course, when X is more efficient, it is more appropriate, but X is even a bit more appropriate than Y when Y would increase efficiency, indicating that the decision maker is not expected to sacrifice for the sake of efficiency when he is already worse off. Hence, the top two rows of the table indicate that efficiency matters, but to a limited extent in the elicited norms when it conflicts with being behind.

Table 2: The difference in norm ratings between X and Y

	Always Ahead	Choice Flips It	Always Behind
Y Efficient ($x_i + x_j < y_i + y_j$)	-0.848	-0.401	0.233
X Efficient ($x_i + x_j \geq y_i + y_j$)	0.350	0.054	0.201
Y More Equal ($\|x_i - x_j\| > \|y_i + y_j\|$)	-0.767	-0.479	0.102
X More Equal ($\|x_i - x_j\| \leq \|y_i + y_j\|$)	0.771	-0.070	0.308
Sacrifice Helps ($y_j > x_j$)	-0.816	-0.385	0.277
Sacrifice Hurts ($y_j \leq x_j$)	0.621	0.154	0.155

This strong shift in norms based on whether one is ahead or behind is perhaps most evident in the bottom part of the table. Choosing Y instead of X is always costly for the dictator and this sacrifice can either help or hurt their partner. When the dictator is ahead, it is much more appropriate to choose Y (compared to X) when doing so would help the other person and much less appropriate to choose Y when it it would hurt them. When the dictator is behind and can sacrifice to help the recipient, it is actually more appropriate to not make the sacrifice ($0.277 > 0$). When the dictator is behind and can give up his own money to harm his partner, it is seen as somewhat more appropriate to not be envious, but the difference shrinks ($0.155 < 0.277$; $p = 0.062$ clustering standard errors by subject). Thus the norms concerning behindness averse actions are relatively weak. In section 3.4, we'll additionally assess the elicited norms by estimating the preference parameters that they would imply if the raters chose purely based on their own normative ratings.[7]

3.2 Estimates for the representative agent

To begin, we estimate the parameters in each of our two models (without and with norms) for the representative agent in both wiggle room conditions, as reported in Table 3. Focusing first on the model without norms and participants in the No Wiggle Room condition, we see that our estimates of α and β are in line with those reported in BFS. The representative agent puts strong positive weight on the other player's payoff when that person is earning less, and lower – but still significantly positive – weight when the other player earns more. For comparison, our estimates are reported alongside the reanalysis of BFS's dictator game data in Table A2.[8] The similarity of the results thus serves as a replication of their work and establishes a similar benchmark for work with our data. In addition, we see that these results are robust to the introduction of moral wiggle room: there are no significant differences in parameter estimates in Table 3 across the two conditions ($p > 0.4$ for α and $p > 0.9$ for β).[9]

Turning to the model with norms in the bottom of Table 3, we see that the representative agent is highly sensitive to norms in both conditions. However, they are even more responsive to norms when there is wiggle room ($p = 0.036$). Most notably, controlling for norms reduces the β estimate in both conditions ($p = 0.058$ for No Wiggle Room and $p < 0.001$ for Wiggle room). This suggests that some of the behavior that was attributed to aheadness aversion in the model without norms may actually be driven by the desire to take appropriate actions. Consistent with our observations that the norms governing actions when the decision maker is ahead are much stronger, accounting for norms affects only the estimate of β and not the α

[7]This table pools across the No Wiggle and Wiggle conditions for readability, but is broken out by condition in Table A1. In twelve of the eighteen cells, there is no significant difference in ratings differences across conditions ($p > 0.10$) In five of the six cases where the ratings varied by condition, the difference was greater under the Wiggle Room condition, indicating that, as expected, the self-interested X was treated as more socially appropriate when there was plausible deniability about implementing it.

[8]Specifically, we estimate our parsimonius (i.e., just α and β) model on their reciprocity-free data to compare with similar estimates using our data.

[9]62 of the 205 people that could wiggle were interrupted by the computer at least once. This accounts for 1% of our observations overall.

estimate.[10] At the same time, it is important to note that norm adherence isn't all that matters: the representative agent still puts positive weight on the payoffs of others even controlling for norms ($p < 0.01$ in all four cases). Assessing the quality of the two specifications using the Akaike information criterion (AIC) indicates that the model with norms fits the data better in both conditions. The Bayesian information criterion (BIC), which penalizes the number of parameters more heavily than the AIC, selects the model with norms only in the Wiggle Room condition.

Table 3: Estimates for Representative Agent

	Model without Norms		
	No Wiggle	Wiggle	z-test: No Wiggle = Wiggle
α	0.076***	0.058***	$p = 0.415$
β	0.336***	0.339***	$p = 0.902$
σ	0.015***	0.015***	$p = 0.268$
Observations	8,730	9,225	
Subjects	194	205	
Log Likelihood	-2,887.75	-3,260.12	
	Model with Norms		
	No Wiggle	Wiggle	z-test: No Wiggle = Wiggle
α	0.078***	0.096***	$p = 0.444$
β	0.263***	0.216***	$p = 0.274$
γ	26.53***	49.30***	$p = 0.036$
σ	0.014***	0.013***	$p = 0.140$
Observations	8,730	9,225	
Subjects	194	205	
Log Likelihood	-2,884.12	-3,245.69	

Notes: p-values calculated using cluster-robust standard errors, clustered by individual.

3.3 Finite mixture model estimates

Rather than assuming that there is a single representative preference type, we next assume that there exist a finite number of types, K, and simultaneously estimate the preference parameters for each type and the classification of individuals into types. To determine the appropriate number of types, we estimate the model for $K = 2, 3, 4$, and 5 and use the NEC to select

[10] If anything, our measured norms indicate that it is slightly more socially appropriate to not help out someone else when already behind, which conflicts with the observed behavior of the representative agent and explains why the α parameter becomes slightly stronger in the model with norms.

the model with the most unambiguous classification into types relative to the change in log likelihood. In all four versions (the two conditions and two specifications), the NEC is lowest when $K = 3$. We therefore report the finite mixture model estimates for each of the three preference types in Table 4.

Table 4: Estimates from Finite Mixture Model

Model without Norms						
	No Wiggle			Wiggle		
	Strong Altruist	Moderate Altruist	Behindness Averse	Strong Altruist	Moderate Altruist	Behindness Averse
Share	0.518***	0.425***	0.057***	0.489***	0.387***	0.123***
α	0.141***	0.044***	-1.65**	0.129***	0.036***	-0.534**
β	0.546***	0.099***	-0.111	0.551***	0.066***	0.313**
σ	0.017***	0.041***	0.003***	0.018***	0.046***	0.004***
Observations	8,730			9,225		
Subjects	194			205		
Log Likelihood	-2,139.75			-2,400.83		
NEC	0.0082			0.0072		
ICL	4,296.16			4,818.49		

Model with Norms						
	No Wiggle			Wiggle		
	Strong Altruist	Norm Follower	Behindness Averse	Strong Altruist	Norm Follower	Behindness Averse
Share	0.519***	0.423***	0.058***	0.492***	0.384***	0.124***
α	0.142***	0.043***	-1.925**	0.164***	0.084***	-0.535*
β	0.531***	-0.078	-0.511	0.453***	-0.177***	-0.08
γ	5.47	54.327***	119.14	43.656***	75.439***	148.393*
σ	0.017***	0.037***	0.002***	0.017***	0.04***	0.004***
Observations	8,730			9,225		
Subjects	194			205		
Log Likelihood	-2,133.93			-2,378.62		
NEC	0.0078			0.0067		
ICL	4,289.55			4,779.00		

Beginning with the model without norms and the No Wiggle Room condition, we identify the same three preference types as BFS: a strong altruist type who puts considerable positive weight on the payoffs of others regardless of being ahead or behind (51.8%), a moderate altruist type who puts weaker, but positive, weight on the payoffs of others (42.5%), and a behindness averse type who puts negative weight on the payoffs of others when behind (5.7%). Our results thus closely align with those of BFS, except that they found somewhat fewer strong altruists (closer to 40%) and more behindness averse types (around 10%).[11]

Again, these results are also largely robust to the inclusion of moral wiggle room, as shown in the top right of Table 4. The one difference is that a significantly higher percentage of subjects are classified as behindness averse (12.3%, up from 5.7%, $p = 0.036$) and this broader type is now also characterized as aheadness averse. Given the preponderance of experimental evidence is that allowing dictator game participants to wiggle out of doing the "right" thing typically makes them appear less altruistic, the impact of our wiggle room manipulation appears consistent with this evidence – overall, we see slightly fewer altruists in the Wiggle Room condition and more inequality averse participants.

Turning to the model with norms, the bottom half of Table 4 indicates that including the social appropriateness of actions in our model does not change the proportions of the population classified as each type (in either wiggle room condition). However, it does reveal distinct motives across types. In the absence of moral wiggle room, strong altruists are not influenced by norms ($p = 0.755$) and accounting for social appropriateness does not influence their preference estimates. In contrast, moderate altruists are highly sensitive to norms ($\gamma = 54.3$, $p = 0.001$), hence their renaming to *norm followers* in Table 4, and once norms are taken into account, it turns out that they don't otherwise care at all about people doing worse than them ($\beta = -0.078$, $p = 0.186$ and the difference in β estimates for the moderate altruists across the two specifications is significant at $p = 0.004$). However, these participants do still place positive weight on their partner's payoff when they are behind, a combination of motivations seemingly more consistent with promoting efficiency. The behindness averse type's estimates are somewhat less stable and precise, but the parameter estimates don't change significantly when we control for norms.

With the introduction of moral wiggle room (lower right of Table 4), we see that norms significantly drive the behavior of all three types ($p < 0.001$ for strong and moderate altruists, $p = 0.078$ for behindness averse types) In addition, the aheadness averse preference parameter collapses to some extent for all three types once we account for norms. Neither the moderate

[11] For comparison, the finite mixture estimates using the BFS data from the dictator (not reciprocity) games are included in Table A3.

altruists nor behindness averse types put positive weight on the payoff of someone doing worse when controlling for norms. Even the strong altruist type, who seemed impervious to norms in the No Wiggle Room condition, exhibits a significant decrease in their β estimate in the model with norms ($p = 0.0099$). However, they still are less motivated by norms than the moderate altruists ($p = 0.0026$) and remain strongly concerned about the outcomes for those who are worse off ($p < 0.001$).

3.4 Is behavior purely driven by subjective perceptions of norms?

We have thus identified different preference types that are differentially driven by normative and outcome-based considerations. An alternative hypothesis is that all participants are primarily norm-driven but they have different subjective perceptions of what normatively appropriate behavior looks like. In this interpretation, the α and β estimates do not purely reflect preferences over outcomes but, rather, misperceptions of the social appropriateness of various actions. To assess this possibility, we use the data from the subjects who rated the normative appropriateness of the choices to simulate what the data might look like if participants were not at all motivated by outcomes and instead driven purely by their own idiosyncratic perceptions of the norm. Specifically, for each pair of allocations considered, we assume that the rater would have chosen the one they personally rated as more socially appropriate. This gives us a simulated data set of 45 choices for each of the 199 participants in the rating portion of the experiment. We then replicate Tables 3 and 4 using these data, as reported in Table A4 and Table A5.

The simulated results differ substantially from our estimates, suggesting that our participants were in fact driven by outcome-based social preferences in addition to the desire to adhere to social norms. In the model without norms, the representative agent is very aheadness averse, with the β parameter close to the maximum of identifiable parameter of 1 that the games were constructed to detect (0.984 in No Wiggle Room and 0.880 in Wiggle Room). In addition, the representative agent is behindness averse ($\alpha < 0$), which is insignificant in No Wiggle Room and highly significantly in Wiggle Room ($p < 0.001$). This thus aligns with the finding in our choice data that behindness aversion is more prevalent when there is Wiggle Room (the difference in α parameters across the simulated data is significant at $p = 0.005$). Most importantly, once we control for norms, none of the outcome-based social preference estimates are statistically significant, with the exception of the marginal significance of α for those in the Wiggle Room condition ($p = 0.086$).

Likewise, the finite mixture model estimates from the simulation differ substantially from our findings, while also providing deeper insight into the heterogeneity in perceived norms than could be gleaned from the aggregate measures in Table 2. Without moral wiggle room, the NEC selects a two-type model with the population roughly equally split between a type that appears strongly altruistic in the model without norms, but is also highly sensitive to norms and no longer aheadness averse when we control for them, and a type that is, in effect, inequality averse (i.e., both aheadness averse and behindness averse). This indicates that the reason social norms have less bite when the decision maker is behind is because the population is roughly split between whether it is more socially appropriate to be altruistic or to reduce disadvantageous inequality, with the former type somewhat more prevalent (representing 56% of our sample). With moral wiggle room, the NEC selects a three-type model. The most common type is both strongly aheadness and behindness averse, with aheadness averse becoming only marginally significant when controlling for norms, while the next most common type is strongly altruistic, with aheadness averse becoming insignificant when controlling for norms.[12]

In short, if our dictators were driven purely by their own perceptions of norms and not at all outcome-driven (and the distribution of these assessments resembled those of our raters), then their choices would generate very different estimates: in particular, we would observe more behindness aversion and find that aheadness aversion is not a strong driver of behavior for the representative agent or any type once we control for norms.

4 Conclusion

Our study contributes by connecting two somewhat independent, but nonetheless influential, strands of the recent literature on social preferences. The first literature catalogues experiments

[12] Recall that the choice data also indicated that more people are classified as behindness averse in the Wiggle Room condition and this result further indicates that more people perceive a norm of inequality aversion when there is wiggle room.

devised to structurally estimate the parameters of an outcome-oriented social preference function based on the weighted average of the decision maker's and a recipient's material outcomes. We contribute here both by replicating recent results using a similar experimental design and by showing the robustness of these estimates with respect to moral wiggle room. The second thread of literature considers how similar choices over material outcomes can be driven by social norms. In this case, we also replicate existing work by documenting the importance of norms for explaining choice behavior.

Our largest contribution, however, lies at the intersection of these two literatures. Not only do our data suggest that both preferences and norms affect choice, we identify the relative weights different preference types place on these motivations and we begin to untangle some of the complicated interactions between the two. For instance, our social appropriateness rating data indicate that the normative implications of being ahead of the recipient are much stronger than the implications of being behind. By estimating a finite mixture model with simulated choice data derived from individual ratings, we find that nearly everyone expects a strong norm of generosity toward someone who earns less ($\beta > 0$) but are about equally split on whether it is socially appropriate to be altruistic or envious toward someone who earns more ($\alpha \lesseqgtr 0$). When the game is played, however, decision makers tend to be more concerned with efficiency than this norm would require. As a result, when controlling for norms in our estimates, we find that this norm absorbs much of the variation previously attributed to the moderate altruist's willingness to help recipients who earn less but not the variation to helping those who already have an advantage. In the end, this means that the previous categorization of players into strong altruists, moderate altruists, and the behindness averse is actually better classified as strong altruists who care little for norms, efficiency-minded norm-followers, and the behindness averse.

References

Andreoni, J. and B. D. Bernheim (2009): "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects," *Econometrica*, 77, 1607–1636.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

Bardsley, N. (2008): "Dictator game giving: altruism or artefact?" *Experimental economics*, 11, 122–133.

Bardsley, N. and P. G. Moffatt (2007): "The experimetrics of public goods: Inferring motivations from contributions," *Theory and Decision*, 62, 161–193.

Bellemare, C., S. Kröger, and A. Van Soest (2008): "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities," *Econometrica*, 76, 815–839.

Bellemare, C., S. Kröger, and A. van Soest (2011): "Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool," *Journal of Economic Behavior & Organization*, 78, 349–365.

Bicchieri, C., E. Dimant, and S. Sonderegger (2020): "It's not a lie if you believe the norm does not apply: conditional norm-following with strategic beliefs," *Available at SSRN 3326146*.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90, 166–193.

Breitmoser, Y. (2013): "Estimation of social preferences in generalized dictator games," *Economics Letters*, 121, 192–197.

Bruhin, A., E. Fehr, and D. Schunk (2019): "The many faces of human sociality: Uncovering the distribution and stability of social preferences," *Journal of the European Economic Association*, 17, 1025–1069.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *The Quarterly Journal of Economics*, 117, 817–869.

Chen, Y. and S. X. Li (2009): "Group identity and social preferences," *American Economic Review*, 99, 431–57.

Conte, A. and M. V. Levati (2014): "Use of data on planned contributions and stated beliefs in the measurement of social preferences," *Theory and Decision*, 76, 201–223.

Conte, A. and P. G. Moffatt (2014): "The econometric modelling of social preferences," *Theory and Decision*, 76, 119–145.

Dana, J., D. M. Cain, and R. M. Dawes (2006): "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and human decision Processes*, 100, 193–201.

Dana, J., R. A. Weber, and J. X. Kuang (2007): "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 33, 67–80.

Engel, C. (2011): "Dictator games: A meta study," *Experimental Economics*, 14, 583–610.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits (2015): "The distributional preferences of an elite," *Science*, 349, aab0096.

Fisman, R., S. Kariv, and D. Markovits (2007): "Individual preferences for giving," *American Economic Review*, 97, 1858–1876.

Grossman, Z. and J. J. Van Der Weele (2017): "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 15, 173–217.

Iriberri, N. and P. Rey-Biel (2011): "The role of role uncertainty in modified dictator games," *Experimental Economics*, 14, 160–180.

——— (2013): "Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do?" *Quantitative Economics*, 4, 515–547.

Kimbrough, E. O. and A. Vostroknutov (2016): "Norms make preferences social," *Journal of the European Economic Association*, 14, 608–638.

Krupka, E. L. and R. A. Weber (2013): "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11, 495–524.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012): "Sorting in experiments with application to social preferences," *American Economic Journal: Applied Economics*, 4, 136–63.

List, J. A. (2007): "On the interpretation of giving in dictator games," *Journal of Political Economy*, 115, 482–493.

McFadden, D. (1981): "Econometric models of probabilistic choice," *Structural analysis of discrete data with econometric applications*, 198272.

Nunnari, S. and M. Pozzi (2022): "Meta-Analysis of inequality aversion estimates," *Working Paper*.

Van der Weele, J. J., J. Kulisa, M. Kosfeld, and G. Friebel (2014): "Resisting moral wiggle room: how robust is reciprocal behavior?" *American Economic Journal: Microeconomics*, 6, 256–64.

Van Leeuwen, B. and I. Alger (2019): "Estimating social preferences and kantian morality in strategic interactions," *TSE Working Paper*.

5 Appendix

Table A1: The difference in norm ratings between X and Y broken down by treatment

	No Wiggle			Wiggle		
	Always Ahead	Choice Flips It	Always Behind	Always Ahead	Choice Flips It	Always Behind
Y Efficient ($x_i + x_j < y_i + y_j$)	-.872	-.487	.143	-.824	-.314	.323
X Efficient ($x_i + x_j \geq y_i + y_j$)	.351	.0745	.18	.350	.033	.223
Y More Equal ($\|x_i - x_j\| > \|y_i + y_j\|$)	-.786	-.547	.121	-.747	-.411	.082
X More Equal ($\|x_i - x_j\| \leq \|y_i + y_j\|$)	.776	-.081	.208	.765	-.059	.409
Sacrifice Helps ($y_j > x_j$)	-.84	-.483	.179	-.791	-.286	.376
Sacrifice Hurts ($y_j \leq x_j$)	.628	.212	.158	.614	.095	.152

Table A2: Estimates for Representative Agent (No Norms) with BFS Comparison

	BFS Session 1	BFS Session 2	No Wiggle	Wiggle	p-value (No Wiggle = Wiggle)
α	0.095***	0.093***	0.076***	0.058***	0.415
β	0.271***	0.240***	0.336***	0.339***	0.902
σ	0.015***	0.019***	0.015***	0.015***	0.268
Observations	18,720	18,720	8,730	9,225	
Subjects	160	160	194	205	

Table A3: BFS Estimates from Finite Mixture Model without Reciprocity

BFS without Reciprocity						
	Session 1			Session 2		
	Strong Altruist	Moderate Altruist	Behindness Averse	Strong Altruist	Moderate Altruist	Behindness Averse
Share	0.408***	0.473***	0.119***	0.366***	0.534***	0.1***
α	0.196***	0.054***	-0.406***	0.202***	0.051***	-0.342***
β	0.494***	0.12***	-0.117	0.494***	0.085***	-0.062
σ	0.017***	0.031***	0.008***	0.019***	0.05***	0.015***
Observations	18,720			18,720		
Subjects	160			160		
Log Likelihood	-4,319.95			-3,241.11		

Table A4: Simulated Choice Data Based on Norm Ratings: Estimates for Representative Agent

	Model without Norms		
	No Wiggle	Wiggle	z-test: No Wiggle = Wiggle
α	-0.064	-0.287***	$p = 0.005$
β	0.984***	0.88***	$p = 0.168$
σ	0.007***	0.008***	$p = 0.047$
Observations	4500	4,455	
Subjects	100	99	
Log Likelihood	-2,107.32	-1,972.50	
	Model with Norms		
	No Wiggle	Wiggle	z-test: No Wiggle = Wiggle
α	-0.086	-0.145*	$p = 0.634$
β	0.048	0.098	$p = 0.599$
γ	672.3***	615.96***	$p = 0.509$
σ	0.005***	0.005***	$p = 0.064$
Observations	4500	4,455	
Subjects	100	99	
Log Likelihood	-1,948.89	-1,808.57	

Notes: p-values calculated using cluster-robust standard errors, clustered by individual.

Table A5: Simulated Choice Data Based on Norm Ratings: Estimates from Finite Mixture Model

	Model without Norms				
	No Wiggle		Wiggle		
	Type 1	Type 2	Type 1	Type 2	Type 3
Share	0.559***	0.441***	0.593***	0.27***	0.137***
α	0.289***	-0.654***	-0.759***	0.223***	-0.021
β	1.047***	0.88***	1.072***	1.017***	0.185**
σ	0.007***	0.009***	0.008***	0.008***	0.011***
Observations	4,500		4,455		
Subjects	100		99		
	Model with Norms				
	No Wiggle		Wiggle		
	Type 1	Type 2	Type 1	Type 2	Type 3
Share	0.558***	0.442***	0.64***	0.265***	0.095***
α	0.528***	-0.928***	-0.724***	0.713***	0.163*
β	-0.054	0.168*	0.195*	0.155	-0.348***
γ	820.18***	491.806***	662.085***	723.068***	182.81**
σ	0.004***	0.007***	0.007***	0.006***	0.008***
Observations	4,500		4,455		
Subjects	100		99		