Modeling choice and estimating utility in simple experimental games

Breitmoser, Yves

EUV Frankfurt (Oder)

21 March 2012

Modeling choice and estimating utility in simple experimental games

Yves Breitmoser[*]
EUV Frankfurt (Oder)

March 21, 2012

Abstract

Current experimental research seeks to estimate shape and parameterization of utility functions. The underlying experimental games tend to be "simple" in that best responses are salient and individual choices consistent, but the analyses arrive at diverging results. I analyze how precision of estimation and robustness of results depend on the choice models used in the analysis. Analyzing dictator and public goods games, I find that regression models overfit drastically when choices exhibit high precision (dictator games), that structural models with simple error structure (normal or extreme value) do not fit the curvature, and that random behavior and random taste models do not identify social motives robustly. The choice process is captured well through the random utility model for ordered alternatives ("ordered GEV", Small, 1987).

JEL–Codes: C50, C72, D64
Keywords: choice models, utility function, estimation, cross validation, dictator game, public goods game

[*] I thank Friedel Bolle, and participants of IMEBE 2011 for many helpful comments, and James Andreoni and John Miller for sharing their data. Address: Europa-Universität Viadrina, Postfach 1786, 15207 Frankfurt(Oder), Germany, email: breitmoser@euv-frankfurt-o.de, Telephone/Fax: +3355534 2291/2390.

1 Introduction

Experimental research in economics has led to a wide range of nowadays consensual insights. For example, experimental subjects have social preferences (Rabin, 1993, Levine, 1998, and subsequent work), subjects do not play strict best responses with respect to any utility function (Rosenthal, 1989, and McKelvey and Palfrey, 1995), but choices are fairly consistent at the individual level (e.g. Andreoni, 1995a).[1] Two influential papers, Andreoni and Miller (2002) and Goeree et al. (2002), then analyzed individual choice patterns in dictator game and public goods games, respectively, for varying tax/subsidy rates and endowments. Their results confirmed the hypothesis that both social preferences and noisiness of choices shape individual decision making—even in simple donation games, where best responses seem salient.

Current quantitative research therefore estimates social motives in relation to econometric models specifying noisiness of choice. Such "structural models" come in three flavors, essentially, and they differ with respect to the location of noise in the choice process. In *random behavior* models (e.g. Fisman et al., 2007), subjects deviate stochastically from their individually optimal response, i.e. noise affects choice after the determination on one's best response. In *random taste* models (e.g. Cox et al., 2007), the altruism coefficient fluctuates randomly, i.e. noise sets in prior to the determination of the best response, and in *random utility* models (e.g. Cappelen et al., 2007), the utilities of all options are perturbed randomly (usually i.i.d.).

Conte and Moffatt (2009, 2010) and Cappelen et al. (2010a) highlight that quantitative results such as the estimated population weights depend on the assumed structure of choice. Arguably, such dependence contributes substantially to the disagreement displayed in the current literature on the social motives underlying pro-social behavior, and thus it puts the underlying, but still unanswered question of Hey (2005) and Loomes (2005) back onto the agenda: How should choice be modeled to robustly identify social motives? The existing literature lacks guidance on this fundamental question, as Card et al. (2011) discuss in detail.

In the present paper, I analyze the choice structure in dictator and public goods games by revisiting the data of Andreoni and Miller (2002) and Goeree et al. (2002). The objective is to understand which components of choice models improve precision

[1] There is further experimental research, e.g. on how choice patterns depend on circumstances (Forsythe et al., 1994; Hoffman et al., 1994; Andreoni, 1995b). See Camerer (2003) for a survey.

and robustness in estimating utility functions.[2] The main results are that structural models avoid significant overfitting,[3] that random utility models yield qualitative robustness of identified motives across treatments, and that generalized errors to adjust kurtosis improve both descriptive and predictive adequacy highly significantly. The ordered GEV model (Small, 1987) has all three of these attributes, to my knowledge currently exclusively, and indeed, its quantitative estimates are precise in-sample, it exhibits insignificant overfitting, and the identified "social motives" (model components) are qualitatively and quantitatively robust across treatments. Guided by the general results, however, additional models may be developed in the future.

In addition, I analyze the reliability of regression models in this context, which relates to a second current debate. The idea is that if one is interested in quantitative estimates of treatment effects, as opposed to being interested in the underlying utility functions as such, then one might circumvent the estimation of "deep" utility parameters and use regression modeling. Proponents of structural modeling argue that the structural constraints reduce overfitting and thereby increase the quantitative robustness of results in relation to say linear regression models (see e.g. Keane, 2010a,b, Rust, 2010, and Nevo and Whinston, 2010). Proponents of non-structural modeling argue that the assumption choice probabilities exactly take the say multinomial logit form is ad hoc and invalid itself.

This debate, too, is one of robustness, as it concerns predictive adequacy with respect to "nearby" treatment conditions. If estimates do not allow for nearby intra- or extrapolation, they give faulty description of the underlying interdependence and cannot be considered robust. Quantitative analyses of such robustness in game-theoretic contexts hardly exist, however. Existing research on predictive robustness has largely focused on decision making under risk and uncertainty (see e.g. Wilcox, 2008, 2011, and Hey et al., 2010). Two exceptions are Cappelen et al. (2011) and Blanco et al. (2011), but the present paper is the first comparative analysis (to my knowledge).

I find that regression models overfit highly significantly if subjects choose with high precision, which applies to dictator games in my analysis, and in this case, their estimates are unreliable quantifications of the treatment effects and of the conditional

[2] Robustness and predictive adequacy will be evaluated by "k-fold cross validation" across treatments: The model is fit to a subset of the treatments, evaluated on the remaining treatments, and the data set is rotated such that all observations are used exactly once in the evaluation stage.

[3] The difference between a model's descriptive and predictive fits is used to measure its tendency to overfit. Overfitting models implies underestimation of standard errors and possibly biased estimates.

choice probabilities. If subjects choose with low precision, as in public goods games in my analysis,[4] then the linear approximations in regression models suffice in the sense that they fit as well as structural models. Even in the latter case, however, regression models do not yield robust identification of components, and including interaction terms increases overfitting in either case. In contrast, structural assumptions prevent models from overfitting in either case, and this holds for all choice specifications in my analysis (the choice specification affects the absolute fit, of course). Thus, I conclude that structural models, and ordered GEV in particular, are preferable to regression models in analyses of social donations.

To outline the intuition, by acknowledging the ordering of alternatives, ordered GEV can represent probability distributions with thin peaks around the utility maximizer (excess kurtosis), which prevail in high precision games, or flat peaks around the maximizer (platykurtosis), which prevail in low precision games. Such adjustment of kurtosis is necessary to obtain efficient estimates and can be captured similarly with generalized normal errors in random behavior and random taste models—but it is largely absent in the current literature (an exception is Cox et al., 2007).

Random taste models fit about as well as ordered GEV in the high-precision dictator games, but they fail to explain choice in public goods games (low-precision choices seem too volatile to be rationalized effectively through "taste" variation). Regression and random behavior models fit about as well in public goods games, but they either fit poorly in the high-precision dictator games (regression) or the identified components are not robust (random behavior). Thus, explaining noisiness of choice as purely statistical "random behavior" proves inferior to relating it to the respective utility differences as in random utility models. Finally, the multinomial logit model of random utility does not allow adjustment of kurtosis, which implies weak levels of goodness-of-fit, but by virtue of being a random utility model, it achieves qualitatively robust identification of components. Thus, all standard models have certain strengths and weaknesses, and in relation to them, random utility with generalized errors (ordered GEV) shares the strengths and avoids most weaknesses.

Section 2 reviews the dictator game experiment of Andreoni and Miller (2002) and the public goods experiment of Goeree et al. (2002). Section 3 describes the

[4] The two experimental games are both games of social donations, but they differ with respect to their strategic complexity, and observed choices therefore have either high or low precision. This can be made quantitatively precise using the hypothetical benchmark model predicting the observed relative frequencies. Its pseudo-R^2 are .579 and .245 in dictator and public goods games, respectively.

currently leading approaches to model choice in this context. Section 4 discusses the independence of irrelevant alternatives (IIA) assumption and the ordered GEV model. Section 5 describes the econometric approach to the analysis. Sections 6 and 7 present and discuss the results for dictator games and public goods games, respectively. Section 8 concludes. The supplementary material contains all parameter estimates, plots illustrating the results, and further analysis illustrating their robustness.

2 Games and data

Dictator games and public goods games are frequently analyzed experimental games and represent the standard frameworks to investigate "social donations." The basic difference between these games is that public goods games have more than one active player. In applications, it may be difficult to distinguish between dictator games and public goods games, however. A donation that is non-strategic in the eye of one person may well be a strategic contribution to a public good (the welfare of the needy) in the eyes of another, i.e. a dictator game can turn into a public goods game depending on who is playing it. In turn, if utilities are linear in the opponents' payoffs, or independent of them, then optimal contributions to (linear) public goods are independent of the opponents' contributions. In such cases, a public goods game degenerates into n independent dictator games. Due these cross-relations between dictator and public goods games, I require a model of social donations to be valid in each of these games.

The data sets of Andreoni and Miller (2002) and Goeree et al. (2002) have been selected for the analysis for three reasons. First, both of these experiments have been designed to evaluate consistency of individual choice, which implies that they are suited to identify the locus of noise. Second, in both experiments, each subject has to make several decisions (eight and ten, respectively) for varying tax/subsidy rates and endowments. This allows me to separate noisy donations from altruistic donations, and thus to identify choice parameters and utility parameters econometrically. This is a main prerequisite for predictive accuracy. Finally, the choice sets are discrete in both experiments and of similar cardinality (ranging from 25 to 100 choices), which suggests that the choice structure may indeed be similar in both cases.[5]

[5] Besides, let me also note that these data sets are made readily available by the authors, either on the Internet (Andreoni and Miller, 2002) or as an appendix (Goeree et al., 2002).

Table 1: Treatment parameters

(a) Dictator game parameters for payoff functions Eq. (1)

Endowment B	40	40	60	60	75	75	60	100
Hold value τ_1	3	1	2	1	2	1	1	1
Pass value τ_2	1	3	1	2	1	2	1	1

(b) Public goods game parameters for payoff functions Eq. (4)

Group size $\lvert N\rvert$	4	2	4	4	2	4	2	2	4	2
Internal return τ_I	4	4	4	2	4	4	2	4	2	4
External return τ_E	2	4	6	2	6	4	6	2	6	12

Dictator games

By Andreoni and Miller's experimental design, the dictator (player 1) can give between 0 and B tokens to player 2, and each token has value τ_1 for player 1 and τ_2 for player 2. Formally, 1's choice set is $S_1 = \{0,1,\dots,B\}$, and the two players' payoffs are, for all $s \in S_1$,

$$\pi_1(s) = \tau_1(B-s) \qquad \text{and} \qquad \pi_2(s) = \tau_2 s. \tag{1}$$

The three parameters τ_1, τ_2, and B vary between treatments,[6] and overall, the 176 subjects of Andreoni and Miller made eight choices each.[7] Table 1a lists the treatment parameters and Figure 1a reviews the distributions of observations. The data exhibits typical characteristics of dictator games. For example, the donations are fairly moderate overall, they are decreasing in τ_1 and increasing in τ_2. A tobit regression of donations on (B,τ_1,τ_2) yields

$$s_1 = \underset{(8.86)}{-8.172} + \underset{(0.066)}{0.254} \cdot B - \underset{(1.871)}{4.348} \cdot \tau_1 + \underset{(1.838)}{4.878} \cdot \tau_2 + \varepsilon \tag{2}$$

where ε has standard deviation $\hat{\sigma} = 27.645$ (the standard errors are provided in parentheses). That is, the donation to player 2 is indeed increasing in the budget and in the donation's value τ_2 for player 2, and it is decreasing in the costs τ_1 for player

[6] I use the term "treatment" to refer to (within subject) variation of the economic environment.

[7] In one session, the subjects had to make three additional choices. I discard these observations, as they are available only for a fifth of all subjects.

1. The regression estimates suggest that the average donation falls by about 4.3 per unit increase of τ_1 and that it increases by about 4.9 on average per unit increase of τ_2. The reliability of such extrapolations is questionable, however, as it requires the underlying model, the tobit model in this case, to have predictive accuracy.

In their analysis, Andreoni and Miller identified subjects with Cobb-Douglas, Leontief, and linear utility functions, and with high or low precision in maximizing utility. These utility functions are special cases of CES utilities

$$u_i(\pi_i,\pi_j) = \Big((1-\alpha)\cdot(1+\pi_i)^{\beta} + \alpha\cdot(1+\pi_j)^{\beta}\Big)^{1/\beta}, \tag{3}$$

where (π_i,π_j) denotes the payoff profile. Cobb-Douglas obtains for $\beta \to 0$, Leontief for $\beta \to -\infty$, and linearity for $\beta = 1$. Similar CES utility functions have been used in most other analyses of dictator games (e.g. Fisman et al., 2007, Cox et al., 2007, Cappelen et al., 2007, and Conte and Moffatt, 2009), and therefore my analysis will be based on CES utilities, too.[8]

Public goods games

The experimental design of Goeree et al. (2002, 32 subjects and 10 treatments) varies the group size n of players consuming the public good as well as external returns τ_E and internal returns τ_I of individual contributions. The costs $\tau_K = 5$ of contributions were held constant and the choice sets are $S_i = \{0,\dots,25\}$ for all players $i \in N$. The payoff of $i \in N$ is, for all $s \in \times_{i\in N} S_i$,

$$\pi_i(s) = \tau_K\cdot(25-s_i) + \tau_I s_i + \tau_E \sum_{j\neq i} s_j. \tag{4}$$

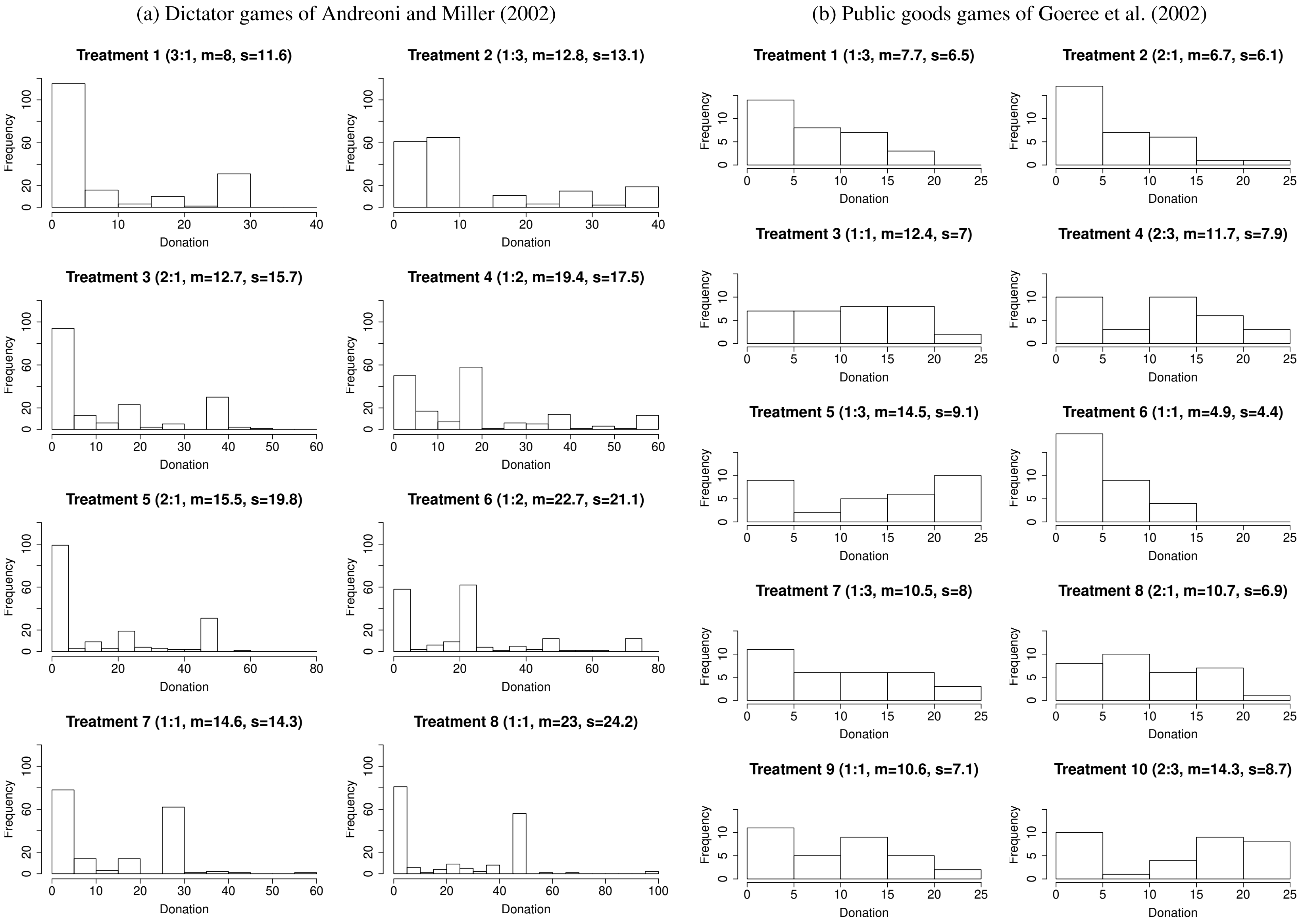
Table 1b and Figure 1b provide overviews of the treatment parameters and the resulting choices. Again, the basic structure of contributions is fairly standard. For example, tobit regressing contributions s_i on the treatment variables (N,τ_I,τ_E) yields

$$s_i = \underset{(2.963)}{-4.721} + \underset{(0.515)}{1.068}\cdot N + \underset{(0.5308)}{2.231}\cdot\tau_I + \underset{(0.1765)}{0.8068}\cdot\tau_E + \varepsilon \tag{5}$$

[8] I use $u_i = -(\text{abs}(\dots))^{1/\beta}$ in case the base in Eq. (3) is negative.

Figure 1: Experimental observations

Note: For each treatment, transfer ratio (τ_1/τ_2 or τ_I/τ_E), as well as mean m and standard deviation s of choices are given in parentheses.



(a) Dictator games of Andreoni and Miller (2002)

(b) Public goods games of Goeree et al. (2002)

where ε has standard deviation $\hat{\sigma} = 8.398$. Thus, contributions are increasing in all treatment variables, internal return, external return, and number of recipients, while the internal return is of higher quantitative relevance than the external return (although the external return is multiplied by $N-1$). In their analysis, Goeree et al. identified players with linear and Cobb-Douglas utility functions, which further underlines the comparability with the data of Andreoni and Miller (2002), as those again are special cases of CES utilities. However, other studies additionally identified "conditional cooperators" (Keser and van Winden, 2000; Fischbacher et al., 2001). Such players wish to contribute about as much as the other players contribute, i.e. their best response functions are increasing with slope close to 1 in the opponents' contributions. In the present case, conditional cooperation can be modeled using Leontief preferences, and besides linear and Cobb-Douglas utilities, Leontief preferences are special cases of the n-player CES aggregator

$$u_i = \left((1-\alpha)\pi_i^\beta + \tfrac{\alpha}{|N|-1}\textstyle\sum_{j\neq i}\pi_j^\beta\right)^{1/\beta}. \tag{6}$$

Aside from the generalization of CES utilities to n-players, the main difference to the dictator game analysis will be the requirement that donations must form a strategic equilibrium in public goods games. The usual approach in (experimental) game theory is to assume that players respond to the distribution of their opponents' choices (following McKelvey and Palfrey, 1995, and Turocy, 2005), about which they have rational expectations. As for public goods games, however, Goeree et al. (2002, Footnotes 20,21) propose that players respond to the expected contributions of the opponents, rather than their full distribution. This approach separates social preferences and risk aversion, as players with non-linear utilities otherwise are risk averse. Similarly, empirical studies separate social motives and risk aversion by assuming that players respond to the actual contributions of others (e.g. of their reference group, as in Andreoni and Scholz, 1998, or the government, see Payne, 1998). For the sake of comparability with these studies, I adopt Goeree et al.'s approach in the following.

3 Current approaches in modeling social donations

This section reviews the models currently used in the literature. Empirical studies on charitable donations generally use regression models (for a review, see Schokkaert, 2006), while experimental studies alternatively use structural models to estimate utility functions (e.g. Cappelen et al., 2007, Cox et al., 2007, and Fisman et al., 2007).

Atheoretic regression Regression models are commonly used in analyses of dictator donations, and some regression models are atheoretic in that they are based on functional forms that are not derived from game-theoretic primitives such as preference orderings and utility maximization. The model considered here regresses donations on treatment parameters (i.e. budget and transfer rates), which also constitutes the standard approach in empirical analyses of charitable donations (see e.g. Auten et al., 2002, and particularly recently, Bakija and Heim, 2011). Since donations in my analysis are discrete, an interval regression with donation s_i as dependent variable is required.

$$s_i = \begin{cases} 0, & \text{if } 0.5 > \hat{s}_i \\ 1, & \text{if } 0.5 \leq \hat{s}_i < 1.5 \\ 2, & \text{if } 1.5 \leq \hat{s}_i < 2.5 \\ \vdots & \\ B, & \text{if } s_i \geq B - 0.5 \end{cases} \qquad \text{with} \quad \hat{s}_i = \alpha + \beta_1 B + \beta_2 \tau_1 + \beta_3 \tau_2 + \varepsilon, \qquad (7)$$

and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. As a second, *extended regression* model, I will consider the model that additionally contains the first-order interaction terms between the treatment variables (B, τ_1, τ_2), i.e. $B \times \tau_1$, $B \times \tau_2$, and $\tau_1 \times \tau_2$. Such interactions are commonly considered in experimental analyses, too.

Structural regression and random behavior The basic idea of random behavior models is that players have deterministic preferences and constant best responses, but they deviate randomly when making the donation. Random behavior models have been used for example by Fisman et al. (2007) and Conte and Moffatt (2010). A similar model structure underlies least squares approaches, which also is used frequently in the literature. Least squares relaxes the distributional assumptions on the the error term, but it cannot be used here, as predicting the distribution of donations is impossible without specifying the error distribution.

Formally, let $u(s|\alpha, \beta)$ denote player 1's utility from donating $s \in S_1$ in a dictator game, and define $\text{BR}(\alpha, \beta) \in \arg\max_{s \in S_1} u(s|\alpha, \beta)$ as the utility maximizing donation if the utility parameters are (α,β) in Eq. (3). Structural regression models are non-linear regression models with the best response function as the deterministic component. For the reasons discussed above, I use an interval regression with latent variable

$$\hat{s}_i = \mathrm{BR}(\alpha,\beta)+\varepsilon. \tag{8}$$

I distinguish two such models. In the *structural regression* model, ε has normal distribution, and in the *random behavior* model, ε has a generalized normal distribution (the exponential power distribution). The former is a fairly standard non-linear model. I call it “structural” as the deterministic component is not a general non-linear function, but the best response. The second model is called “random behavior” to emphasize the behavioral component implied by the generalized error distribution. Their joint analysis will enable us to look at the relevance of the error specification.

Random taste An alternative approach toward explaining noisiness of choice is to allow for one’s interest in the opponent’s well-being to fluctuate randomly (Cox et al., 2007). That is, the altruism coefficient α or “taste” is random. The key limitation of assuming random taste is that this limits choices to those rationalizable by variations of α. This limitation does not seem too severe in dictator or public goods games, where variations of α seem sufficient, but in general the random taste assumption does not suffice. Formally, the donation in the random taste model is

$$s_i = \mathrm{BR}(\alpha+\varepsilon,\beta), \tag{9}$$

where ε again has a generalized normal distribution. The specification used here follows Cox et al. (2007): $\alpha' := \alpha/(1-\alpha)$ has exponential power distribution with mean m, scale s, shape ρ, i.e. density $f(\alpha') = \rho \exp\left\{(|\alpha'-m|/s)^{\rho}\right\}/2s\Gamma(1/\rho)$. The implied probability that i chooses an action $s_i' \leq s_i$ equates with $F(\alpha^*)$ where α^* solves $u_i(s_i|\alpha^*) = u_i(s_i+1|\alpha^*)$. See Cox et al. (2007, Appendix B) for illustrations.

Random utility assuming IIA The random utility model satisfying independence of irrelevant alternatives (IIA), *multinomial logit*, is used very frequently in structural analyses of experimental data (see Cappelen et al., 2007, and the references in the next section). Here players maximize utility, while the utilities of the various options fluctuate randomly. Hence players tend to deviate from the unperturbed optimal choice but choose "better" options with higher probability. Similarly to random behavior, randomness of utilities may explain all deviations from utility maximization, but now the probability of deviating to a given alternative depends on the loss that it induces rather than its distance in the choice set.

Formally, the players are assumed to maximize $\tilde{u}(s) = \lambda u(s|\alpha,\beta) + \varepsilon_s$ over $s \in S_1$, with $\lambda \geq 0$ and ε_s being i.i.d. extreme value distributed (for all options $s \in S_1$). This distribution of ε_s yields choice probabilities with the multinomial logit form and implies IIA (Luce, 1959),

$$\forall s \in S_1 : \qquad \Pr(s) = e^{\lambda u(s|\alpha,\beta)} / \sum_{s' \in S_1} e^{\lambda u(s'|\alpha,\beta)}. \tag{10}$$

4 Random utility without IIA

Random utility models as they are used in experimental game theory generally assume i.i.d. random components ε_s. Examples include Anderson et al. (2001, 2002), Kübler and Weizsäcker (2004, 2005), and Costa-Gomes et al. (2009).[9] The independence of the utility perturbations induces independence of irrelevant alternatives (IIA): when a choice $s \in S_1$ is eliminated from the choice set S_1, the relative odds between any remaining choices s' and s'' are unaffected. IIA has been criticized frequently in the econometric literature, however. Tversky (1972) writes "the addition of an alternative to an ordered set 'hurts' alternatives that are similar to the added alternative more than those that are dissimilar to it," and a similar point is made with the red bus-blue bus example of McFadden (1973).[10] General discussions of IIA can also be found in McFadden (1976) and Samuelson (1985).

If choice sets are ordered, then similarity in Tversky's sense may relate to the proximity of choices. This idea has been introduced by Small (1987, 1994).[11] Small's

[9] This is different in other branches of economics, in particular in modeling demand functions e.g. for residential location (McFadden, 1978), the use of telephone services (Train et al., 1987; Lee, 1999), multi-product firms (Anderson and De Palma, 1992), for urban travel (see e.g. Train, 2003, for a review), and also in modeling voting behavior (Whitten and Palmer, 1996).

[10] One can commute either by car or by bus, and initially, 50% of the commuters choose either option. Now a second bus is introduced which is in all ways equivalent to the first one (and capacity constraints were never an issue). Intuitively, 50% of the commuters still choose to go by car, but under IIA, only a third of them is predicted to do so.

[11] The supplementary material contains further results for an alternative models relaxing IIA, a

"ordered GEV" model essentially introduces a parameter $\rho \in (0,1]$ to capture the correlation of utility perturbations between proximate choices and "weights" to capture how correlation ceases as options become more distant. The way the model is usually defined, stochastic independence (i.e. multinomial logit) obtains for $\rho = 1$ and perfect correlation obtains for $\rho \to 0$.

To define the model, let $M \in \mathbb{N}_0$ denote a bandwidth parameter, and define $\rho > 0$ and weights $w_m \geq 0$ for all $m = 0, \dots, M$ such that $\sum_{m=0}^{M} w_m = 1$. In the analysis, I use Gaussian weights.[12] The ordered GEV choice probabilities are

$$\omega(s) = \sum_{r=s}^{s+M} \frac{w_{r-s} \exp\left\{\lambda u(s|\alpha,\beta)/\rho\right\}}{\exp\{I_r\}} \cdot \frac{\exp\{\rho I_r\}}{\sum_{t=0}^{B+M} \exp\{\rho I_t\}} \tag{11}$$

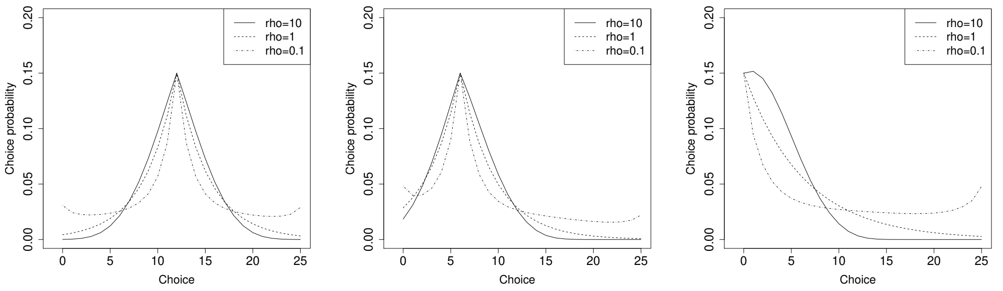
with inclusive value $I_r = \ln \sum_{s' \in B_r} w_{r-s'} \exp\left\{\lambda u(s'|\alpha,\beta)/\rho\right\}$ for all $r \in \{0,\dots,B+M\}$ and nests $B_r = \left\{s \in \{0,1,\dots,B\} \mid r - M \leq s \leq r\right\}$. That is, the player first chooses a nest B_r, $r \in \{0,\dots,B+M\}$, and secondly he chooses an option $s \in B_r$ in this nest. The probability of choosing s conditional on having chosen B_r is the first factor in each of the summands of Eq. (11), and the probability of choosing nest B_r in the first place is the second factor above. The products of these probabilities are aggregated over all neighborhoods containing s. Since every strategy belongs to $M+1$ nests, nests are overlapping and the model is "cross-nested" in the sense of Vovsha (1997). Ordered GEV also is a special case of "elimination by aspect" (Tversky, 1972), as illustrated by McFadden (1981, p. 225f).

Figure 2 illustrates the choice pattern implied by varying ρ in the ordered GEV model. In general, $\rho < 1$ implies that options with high base utility attract probability mass away from proximate options. To see this, consider two options with perfectly correlated utility perturbations (i.e. $\rho \to 0$). Now, the option with the higher base utility between the two is preferred to the other one with probability one even after perturbations. The effect is weakened but similar for all $\rho < 1$. Thus, the choice distributions have *leptokurtosis* (excess kurtosis), i.e. narrow peaks and fatter tails than

nested logit model, which further illustrate that models relaxing IIA are not particularly vulnerable to overfitting. This models improve upon multinomial logit as well, but it is not as accurate as accounting for ordering directly (as in ordered GEV) and hence skipped for brevity.

[12] That is, $w_m = f_{\mathcal{N}(M/2,\sigma^2)}(m) / \sum_{m=0}^{M} f_{\mathcal{N}(M/2,\sigma^2)}(m)$ with free parameter σ, $f_{\mathcal{N}(\mu,\sigma^2)}$ denotes the density of the normal distribution with mean $M/2$ and variance σ^2. The theoretical bandwidth is set to be $M = B/2$ rounded up to the nearest even, but it is practically irrelevant, as the estimated σ will be rather low. To improve comparability with the other structural models, which require up to four parameters per component, I assume that σ is constant across components.

Figure 2: Ordered GEV responses for three stylized utility functions



(a) $u(x) = -|x_i - 12|$

(b) $u(x) = -|x_i - 6|$

(c) $u(x) = -|x_i - 0|$

Note: The OGEV responses, see Eq. (11), are given for choices from $\{0, 1, \ldots, 25\}$, with the standard deviation of the Gaussian weights (Fn. 12) being 4, the value of ρ being as defined in the plots, and λ being set such that the choice probability of the option maximizing base utility is 0.15 (to ensure comparability).

they display in the multinomial logit case $\rho = 1$. Such distributions capture choice patterns where the base utility maximizer is chosen with supra-proportional probability but not exclusively. Arguably, such choice patterns can be expected if the utility maximizers are particularly salient. In turn, $\rho > 1$ allows to express *platykurtosis*, i.e. flat peaks and thin tails. Such models are not random utility models in the sense of (McFadden, 1978), as respective choice probabilities cannot be explained by stochastic utility perturbations, but they may fit choice patterns when utility maximizers are not salient.

5 Econometric approach to the analysis

This section describes the general approach to the analysis and the two basic measures for goodness of fit. As indicated before, both Andreoni and Miller (2002) and Goeree et al. (2002) identified subjects with either linear, Cobb-Douglas, or Leontief utility functions, and high or low precision. Subject heterogeneity of such discrete nature is aptly modeled as a finite mixture (McLachlan and Peel, 2000), which in turn is standard practice in experimental analyses. For example Stahl and Wilson (1995) and Kübler and Weizsäcker (2004) model heterogeneity in strategic reasoning as finite mixtures, Harrison and Rutström (2009), Bruhin et al. (2010), and Conte et al. (2011) model choice under risk, Cappelen et al. (2007, 2010b) model dictator donations, and

Bardsley and Moffatt (2007) model public good contributions.

To define the finite mixture model, let K denote the set of subject types ("components") in the population, e.g. $K = \{A, B, C\}$ in a three-component population, let $(\nu_k)_{k \in K}$ denote the component shares, and for all $k \in K$, let P_k denote the parameter profile characterizing component k. Now, if $o_{j,t}$ denotes the tth observation of subject $j \in J$ in the data set and if $\omega(o_{j,t}|P_k)$ is the probability that j chooses $o_{j,t}$ conditional on being of component k, the log-likelihood of the finite mixture model is

$$LL(P, \nu | o) = \sum_{j \in J} \ln \sum_{k \in K} \nu_k L(j,k) \qquad \text{with} \quad L(j,k) = \prod_t \omega(o_{j,t}|P_k). \tag{12}$$

In the models estimated below, members of different components have different utility functions and different precision in maximizing their utilities. In conjunction, these differences encompass the different motivations for their pro-social choices. For this reason, I will occasionally refer to the identified components as the identified *social motives*.

The choice probabilities $\omega(o_{j,t}|P_k)$ needed to compute the likelihood follow immediately from the above definitions. The model parameters are estimated by maximizing the likelihood function, which I maximize jointly over all parameters. This helps to avoid certain issues with two-step estimators (e.g. inconsistency and inefficiency, as discussed by Amemiya, 1978, and Arcidiacono and Jones, 2003) and it allows to extract estimates of the standard errors from the information matrix. I use the global maximizer CMAES for the initial approach toward the maximum (an evolutionary strategy, see Hansen et al., 2003, and Hansen and Kern, 2004), subsequently the derivative-free NEWUOA algorithm (Powell, 2008) to search the wider neighborhood of the initial estimates (NEWUOA is a comparably robust algorithm, see Auger et al., 2009, and Moré and Wild, 2009), and finally a Newton-Raphson algorithm to ensure local convergence. In order to further ensure global convergence, the above procedure had been restarted repeatedly varying the starting values.

Descriptive accuracy (model precision) Assessment of the goodness of fit of econometric models are usually based on measures such as Bayes' information criterion $BIC = -LL + d/2 \cdot \ln O$ with d as the number of parameters of the model and O as the number of independent observations (the number of subjects). The term $d/2 \cdot \ln O$ approximates the correction required to offset the fact that additional parameters necessarily raise the maximum of the likelihood function (for the underlying assumptions, see Schwarz, 1978). In finite mixture models, this correction is insufficient and BIC tends to overestimate the number of components. Biernacki et al. (1999, 2000) propose to use the complete-data likelihood of the mixture model (which additionally accounts for the likelihood of the component membership indicators) instead of the likelihood itself to address this shortcoming. The resulting integrated classification likelihood (ICL) criterion is derived under assumptions otherwise equivalent to those of BIC, and can be approximated by

$$\begin{aligned} & \mathit{ICL\text{-}BIC} = -LL + d/2 \cdot \ln O + \mathrm{En}(\hat{\tau}) \\ & \qquad \text{with } \mathrm{En}(\hat{\tau}) = -\sum_{j \in J} \sum_{k \in K} \hat{\tau}_{jk} \ln \hat{\tau}_{jk} \quad \text{with} \quad \hat{\tau}_{jk} = \frac{\nu_k L(j,k)}{\sum_{k' \in K} \nu_{k'} L(j,k')}, \end{aligned} \tag{13}$$

where $\hat{\tau}_{jk}$ is the posterior probability of j's membership in component k based on the parameter estimates. The "entropy" $\mathrm{En}(\hat{\tau})$ of the classification matrix $\left(\hat{\tau}_{jk}\right)_{j,k}$ penalizes mixture models with poorly separated, thus superfluous components. I will use *ICL-BIC* in order to assess the goodness of fit in-sample. McLachlan and Peel (2000, Chapter 6) discuss *ICL-BIC* and alternative measures in more detail.

Predictive accuracy (model robustness) A complementary approach toward model validation is the assessment of its predictive accuracy, as this specifically measures the degree to which the model captures the actual data generating process. For a general discussion, let me refer to Keane and Wolpin (2007). The predictive accuracy is especially interesting for models of social preferences, as their lack of robustness between studies, and even within subjects (Blanco et al., 2011), is a current topic in experimental analyses. For this reason, I analyze to which degree robustness *across treatments* depends on model choice.[13]

That is, I evaluate predictive accuracy by treatment-based cross validation (Burman, 1989; Zhang, 1993) with non-random holdout samples (Keane and Wolpin, 2007). The data set is partitioned into $K = 4$ "folds" containing two/three treatments each. I use the observations from three folds to estimate the parameters and then

[13] An alternative approach would consider predictions that are both across treatments and across subjects, e.g. across experiments. While this kind of robustness would be very desirable, it is left as further research, as it builds on the results derived below and additionally requires analysts to efficiently capture heterogeneity of subject pools (such as heterogeneity in their composition with respect to social motives and levels of reasoning), which also is a current research topic.

Table 2: Partitioning of the data sets to assess predictive accuracy

(a) Dictator game analysis

Treatments used for …		
Training		Validation
$\{1,2,3,4,5,7\}$	$\rightsquigarrow$	$S_1 = \{6,8\}$
$\{1,2,3,4,6,8\}$	$\rightsquigarrow$	$S_2 = \{5,7\}$
$\{1,3,5,6,7,8\}$	$\rightsquigarrow$	$S_3 = \{2,4\}$
$\{2,4,5,6,7,8\}$	$\rightsquigarrow$	$S_4 = \{1,3\}$

(b) Public goods analysis

Treatments used for …		
Training		Validation
$\{1,2,3,4,5,6,7\}$	$\rightsquigarrow$	$S_1 = \{8,9,10\}$
$\{1,2,3,4,8,9,10\}$	$\rightsquigarrow$	$S_2 = \{5,6,7\}$
$\{1,2,5,6,7,8,9,10\}$	$\rightsquigarrow$	$S_3 = \{3,4\}$
$\{3,4,5,6,7,8,9,10\}$	$\rightsquigarrow$	$S_4 = \{1,2\}$

compute their log-likelihood on the fourth fold to gauge their predictive accuracy. By rotating such that each fold is used exactly once in the validation stage, and aggregating the likelihoods, I obtain the *cross validation-based information criterion* denoted LL_{Out}. Smyth (2000) discusses cross validation in the context of mixture models. Using $\big(P_{(f)}, \nu_{(f)}\big)$ as the parameter estimates if fold S_f ($f = 1, \ldots, 4$) is left out, and $LL(\cdot|S_f)$ as the log-likelihood using these estimates evaluated on fold S_f, this criterion and its in-sample counter-part is

$$LL_{\text{Out}} = -\sum_{f=1}^{4} LL\big(P_{(f)}, \nu_{(f)}|S_f\big), \qquad\qquad LL_{\text{In}} = -\sum_{f=1}^{4} LL\big(P, \nu|S_f\big), \tag{14}$$

with (P, ν) as the whole-sample likelihood maximizers. The folds used in the analysis are defined in Tables 2a and 2b. They are defined such that the respective holdout samples are not “extreme” but different. For example, in the dictator game data, with fold S_3, I use 5 observations per subject for transfer rates $1 : x$, $x \geq 1$, and one observation for $2 : 1$, to predict donations in case of $2 : 1$ and $3 : 1$.

Using cross validation, I evaluate robustness in three (partially) complementary ways. First, *qualitative robustness* obtains if no superfluous components are identified. It is assessed by comparing the number of components identified in the whole sample with the number of components that are robust to cross validation. The former *in-sample optimal* number of components is the number that minimizes *ICL-BIC*, while the latter *out-of-sample optimal* number of components is the smallest number with predictive accuracy LL_{Out} that is not significantly improved upon at $\alpha = 0.1$ (this moderate significance threshold is introduced to mildly penalize non-parsimonious models, similarly to the effect of the BIC correction term in *ICL-BIC*).

Second, *quantitative robustness* is assessed by evaluating the significance of differences between training-sample estimates and whole-sample estimates. The possible approaches to evaluate these differences vary in the data on which the significance of differences is evaluated. I use two such approaches. On the one hand, I evaluate the significance on the data folds left out in the respective training samples. This yields the null hypothesis $LL_{\mathrm{In}} = LL_{\mathrm{Out}}$, which is evaluated in non-nested Vuong tests where $LL_{\mathrm{In}} - LL_{\mathrm{Out}} > 0$ is significant if it exceeds the BIC correction term significantly (adopting the suggestion of Vuong, 1989, Eq. 5.9). As this test evaluates a model's fallacy to overfit on the training samples, I refer to it as *LR test of overfitting*.

On the other hand, I evaluate the differences of whole-sample estimates and training-sample estimates on the whole data set. This tests the robustness of model estimates with respect to extending the data set from the training sample to the whole sample. The respective null hypothesis $LL(P_{(f)}, \nu_{(f)}) = LL(P, \nu)$ is evaluated (for all folds f) in Vuong tests *without* BIC correction, as the parameter spaces have the same dimensionality and the respective data sets have the same number of independent observations. Due to dropping the BIC correction, this is a rather strict test of quantitative robustness. I refer to it as the *LR test of robustness*.

6 Modeling choice in dictator games

In the next two sections, I discuss the properties of the choice models with respect to the dictator and public goods data, respectively. The key information for dictator games is summarized in Tables 3 and 4 below. All parameter estimates and several illustrating plots are provided as supplementary material.

First, Table 3 provides an overview of the in-sample fit for all models with up two seven components. Further components are not required, as the predictive accuracy LL_{out} (which has been estimated simultaneously) peaks for all but the random-utility models at five components or less, and the in-sample fit *ICL-BIC* drops beyond seven components for the random-utility models. This corresponds with Andreoni and Miller (2002), who identified six distinct components in their analysis. Table 3 also lists the results of likelihood ratio tests between the models with the "in-sample optimal" (by *ICL-BIC*) numbers of components. I consider differences that induce p-values less than .01 to be significant. The following summarizes the main observa-

Table 3: In-sample summary for the dictator games

(a) Descriptive adequacy $\mathit{ICL\text{-}BIC} = -LL + d/2 \cdot \ln O + \mathrm{En}(\hat{\tau})$ (less is better)

		Number of components						
	#Pars	One	Two	Three	Four	Five	Six	Seven
Regression	5	4270	3773	3659	3473	3415	3381	**3359**
Extended regression	8	4274	3777	3657	2976	2918	2898	**2879**
Structural regression	3	4341	3407	3097	2969	2899	**2872**	2874
Random behavior	4	4337	3243	3088	2841	2686	**2641**	2646
Random taste	4	4316	3481	3053	2951	2713	**2645**	2645
Random utility	3	4724	3923	3258	3086	3042	3010	**2935**
Ordered GEV	4	4292	3046	2945	2846	2787	2682	**2626**

(b) LR tests on the descriptive adequacy using the in-sample-optimal number of components

		Model 2						
Model 1	CLC	Regr 7	ExtReg 7	StructReg 6	RBehav 6	RTaste 6	RUtil 7	OGEV 7
Regr 7	3253.3		<<<	<<<	<<<	<<<	<<<	<<<
ExtReg 7	2718.72	>>>		=	<<<	<<<	=	<<<
StructReg 6	2812.38	>>>	=		<<<	<<<	=	<<<
RBehav 6	2566.36	>>>	>>>	>>>		=	>>	=
RTaste 6	2570.02	>>>	>>>	>>>	=		>>>	=
RUtil 7	2828.68	>>>	=	=	<<	<<<		<<<
OGEV 7	2520.33	>>>	>>>	>>>	=	=	>>>	
Baseline 1	5814	<<<	<<<	<<<	<<<	<<<	<<<	<<<

Note: "Baseline 1" represents the prediction of uniform randomization, and otherwise, "Model K" represents the model with K components. $CLC = -LL + \mathrm{En}(\hat{\tau})$ is the classification likelihood criterion. The null hypotheses $H_0 : CLC(\mathit{Model\ 1}) = CLC(\mathit{Model\ 2})$ are evaluated in Vuong LR tests accounting for the entropies $\mathrm{En}(\hat{\tau})$, and in non-nested cases also including BIC correction terms (as suggested in Eq. 5.9 of Vuong, 1989). "$\ggg, \gg, >$" indicate rejection of H_0 in favor of Model 1 with $p < .001, .01, .1$ (resp.), "$\lll, \ll, <$" indicate rejection in favor of Model 2, and "$=$" indicate no rejection.

tions from Table 3.

Result 6.1 (Descriptive adequacy)**.** *By their fit to the whole sample, three tiers of models can be distinguished. First, the "advanced" structural models (random behavior, random taste and ordered GEV) are not significantly different at* $\alpha = .01$*, while all three of them fit highly significantly (*$p < .01$*) better than all other models. Second, the basic structural models (random utility and structural regression) and the extended regression model are not significantly different, but all of them improve upon the final model, linear regression.*

I call the three best-fitting models "advanced" because all three of them have a free parameter to adjust the shape of the error distribution. The random behavior and random taste models allow for generalized normal errors, which can have leptokurtic shapes (thin peaks and fat tails) or platykurtic shapes (fat peaks and thin tails), and ordered GEV allows to adjust kurtosis by varying ρ (as shown in Figure 2). Thus, with respect to descriptive adequacy, the error specification dominates the error location in structural models. The estimated components and utility parameters differ between models, however, as shown in Table 5 and discussed shortly in detail, and these differences will affect robustness and predictive accuracy.

Nonetheless, these results illustrate the necessity to capture excess kurtosis in dictator games, and to relax independence of irrelevant alternatives in these cases (note that all three "advanced" models violate IIA). With few exceptions (e.g. Cox et al., 2007), such generalized models are not considered yet. Before discussing these observations in further detail, let us verify their robustness out of sample (Table 4).

Result 6.2 (Predictive accuracy). *Ordered GEV fits significantly better ($p < .01$) than all other models—even if we adjust their numbers of components from the in-sample optimum to the out-of-sample optimum.*[14] *The other two advanced structural models still fit significantly better than random utility with IIA.*

The ordered GEV model of random utility predicts most accurately. In turn, the components identified by the other two "advanced" models, random behavior and random taste, are less robust across treatments. In particular, the numbers of components that remain distinct out-of-sample drop from six to four and five for random behavior and random taste, respectively. Consequently, their out-of-sample fit is worse than ordered GEV's. That is, ordered GEV model does not achieve a significantly better fit in-sample, but it achieves this goodness-of-fit more robustly.

Result 6.3 (Qualitative robustness). *Both random utility models (multinomial logit and ordered GEV) achieve "precise" and "robust" identification of the number of components, i.e. the number of components identified in-sample is higher ($K = 7$) than those of the other structural models, and all identified components are robust to cross validation. None of the alternative structural models yields either precision or robustness in this sense.*

[14] As described above, *out-of-sample optimal* number of components is the smallest number that is not significantly improved upon at $\alpha = 0.1$. Using the in-sample optimal number of components would further strengthen the result in favor of ordered GEV.

Table 4: Summary of the out-of-sample analysis of dictator game models

(a) The out-of-sample log-likelihoods LL_{Out} Eq. (14) and optimal number of components

	Number of components												
Model 1	1		2		3		4		5		6		7
Regr	−4291.01	$\lll$	**− 3977.01**	$=$	−4292.79	$<$	−4250.87	$=$	−4401.51	$\lll$	−4354.85	$\lll$	−4221.45
ExtReg	−4292.88	$\lll$	−3995.63	$=$	−4217.06	$=$	−4215.24	$\lll$	**−3779.5**	$=$	−4327.79	$\ll$	−4272.24
StructReg	−4363.7	$\lll$	−3403.32	$<$	−3346.36	$\lll$	−3140.2	$\ll$	**−3104.82**	$=$	−3109.03	$=$	−3103.44
RBehav	−4386.88	$\lll$	−3279.39	$<$	−3231.46	$\lll$	**−2916.91**	$=$	−2909.39	$=$	−2919	$<$	−2913.21
RTaste	−4474.44	$\lll$	−3525.55	$\lll$	−3199.19	$<$	−3154.33	$\lll$	**−2907.17**	$=$	−2904.05	$=$	−2902.23
RUtil	−4830.86	$\lll$	−4115.95	$\lll$	−3537.32	$\lll$	−3235.43	$<$	−3207.68	$<$	−3192.59	$\ll$	**−3112.67**
OGEV	−4333.35	$\lll$	−3157.81	$=$	−3162.53	$\lll$	−3093.7	$<$	−3047.88	$\lll$	−2872.41	$\lll$	**−2820.16**

Note: $\lll, \ll, <$ indicate rejection of $H_0 : LL_{\text{Out}}(k\ \textit{Components}) = LL_{\text{Out}}(k+1\ \textit{Components})$ with $p < .001, .01, .1$. The log-likelihood LL_{Out} for the *out-of-sample optimal* number of components (smallest number of components that is not significantly improved upon at $\alpha = 0.1$) is set in bold-face type.

(b) Likelihood-ratio tests of LL_{Out} using the out-of-sample optimal number of components

		Model 2						
Model 1	LL	Regr 2	ExtReg 5	StructReg 5	RBehav 4	RTaste 5	RUtil 7	OGEV 7
Regr 2	−3977.01		<<<	<<<	<<<	<<<	<<<	<<<
ExtReg 5	−3779.5	>>>		<<<	<<<	<<<	<<<	<<<
StructReg 5	−3104.82	>>>	>>>		<<<	<<<	=	<<<
RBehav 4	−2916.91	>>>	>>>	>>>		=	>>>	<<<
RTaste 5	−2907.17	>>>	>>>	>>>	=		>>>	<<
RUtil 7	−3112.67	>>>	>>>	=	<<<	<<<		<<<
OGEV 7	−2820.16	>>>	>>>	>>>	>>>	>>	>>>	
Baseline 1	−5814	<<<	<<<	<<<	<<<	<<<	<<<	<<<

Note: "Model *K*" represents the Model with *K* components, where *K* is the smallest number of components that is not significantly improved upon at $\alpha = 0.1$ (see Table 4a). The null hypothesis in the tests is $H_0 : LL_{\text{Out}}(\textit{Model 1}) = LL_{\text{Out}}(\textit{Model 2})$ and evaluated in Vuong tests for non-nested models; $\ggg, \gg, >$ indicate rejection at $p < .001, .01, .1$.

(c) Likelihood-ratio tests on overfitting and robustness of parameter estimates (see Section 5)

	LR tests on overfitting for number of components $K = \ldots$								LR tests on robustness for folds …			
Model 1	1	2	3	4	5	6	7		S_1	S_2	S_3	S_4
Regr	>>>	>	>>>	>>>	>>>	>>>	>>>	Regr 7	>>>	>>>	>>>	>>>
ExtReg	>>	>	>>>	>>>	>>>	>>>	>>>	ExtReg 7	>>>	>>>	>>>	>>>
StructReg	>>>	>>	>>>	=	=	=	=	StructReg 6	=	=	=	=
RBehav	>>>	>>>	=	=	=	=	=	RBehav 6	>>>	>	>	=
RTaste	>>>	>>>	>>	=	=	>	=	RTaste 6	>>	>>	=	>
RUtil	>>>	>>>	>>>	=	=	=	=	RUtil 7	>>	=	=	>>
OGEV	>>>	=	>>>	=	>	=	=	OGEV 7	>	>	>	>

Overfitting: $H_0 : LL_{\text{In}}(\textit{Model 1}) = LL_{\text{Out}}(\textit{Model 1})$ evaluated in one-sided, non-nested Vuong tests (with BIC correction term applied to LL_{in}); $\ggg, \gg, >$ indicate significant overfitting at $p < .001, .01, .1$.
Robustness: $H_0 : LL(P_{(f)}, \nu_{(f)}) = LL(P, \nu)$ is evaluated for all folds $f \in \{1, 2, 3, 4\}$, see above.

The observation that identified motives may become indistinct out of sample would explain the lacking robustness of behavior across games currently discussed in the literature (e.g. Blanco et al., 2011) and the lacking robustness across studies shaping the literature. Apparently, only random utility models with or without IIA yield qualitative robustness in dictator games such as those analyzed here.

To see why, let us next look at Table 5. It describes the identified subject types (using the whole sample) for both in-sample and out-of-sample optimal number of components. In all cases, strictly egoistic and strictly Leontief subjects are identified (components 1 and 2, respectively). These subjects do not deviate from utility maximization, and hence their identification does not depend on the assumed error structure. In the whole sample, both random behavior and random taste identified four additional components containing "noisy" players, but only two or three of them (respectively) are robust to cross validation.

As for random behavior, the components number 4–6 in Table 5 are indistinct out-of-sample, and these components comprise 42% of the subjects. The only substantial difference between random behavior and ordered GEV is the difference in the location of the error. Hence, this location affects the robustness. As an illustration, consider the extreme case that a model is fitted to data for a 3 : 1 transfer ratio, i.e. $\tau_1 : \tau_2$ in Eq. (1), and is used to predict data for a 1 : 1 or 1 : 3 transfer ratio.[15] In case 3 : 1, the subjective value of a token is substantially higher than in case 1 : 1, even for altruistic subjects. As a result, the observed standard deviations differ between treatments (see also Figure 1a). Primarily, the variances of "intermediately precise" players adapt to changes in the transfer ratios. The random behavior model requires this subjective value to not affect the standard deviation of choices, however. Thus, it does not allow robust identification of the intermediately precise components (numbers 4–6), which become indistinct out-of-sample.

To be more specific, Table 4c shows that the random behavior model does not predict behavior in fold S_1 robustly, i.e. in treatments 6 and 8 where observations have the highest variances in absolute terms. Random utility modeling, in contrast, posits that the probability of chosing a given option is inversely related to its opportunity costs—and the opportunity costs depend on the transfer ratio.[16] This reduces the

[15] The in-sample/out-of-sample ratios used in the analysis are far less extreme, see Table 2a, but the tendency is similar.

[16] For a more explicit analysis of such optimization incentives, see Battalio et al. (2001).

Table 5: The components estimated by the “advanced models” for the whole sample, using either the in-sample or the out-of-sample optimal number of components

Model	Components							
	#1		#2		#3		#4	
	#5		#6		#7			
Random behavior								
In-sample optimal number	−0.757	10.193	0.251	−67.791	0.501	0.49	0.347	0.174
	0.238	0	0.142	0	0.175	35.3	0.16	10.9
	0.49	−30.759	0.097	0.296				
	0.123	27.4	0.161	9.1				
Out-of-sample optim. number	−0.89	9.586	0.303	−86.109	0.501	0.488	0.298	0.407
	0.239	0	0.153	0.2	0.191	34.7	0.417	14.5
Random taste								
In-sample optimal number	−0.93	29.431	1.229	−62.287	0.999	0.485	0.5	0.847
	0.238	0	0.154	4.4	0.124	29.3	0.176	32.2
	0.338	−0.035	0.322	0.444				
	0.235	18.8	0.073	8.8				
Out-of-sample optim. number	−0.381	4.927	1.018	−55.579	0.999	0.485	0.5	0.845
	0.238	0	0.154	4.5	0.123	29.2	0.185	32
	0.34	−0.013						
	0.3	18.4						
Ordered GEV								
In + Out	−0.769	−0.045	0.291	−79.613	0.5	0.491	0.331	−100
	0.25	0	0.142	0	0.083	12.7	0.129	22.2
	0.237	−0.374	0.361	1.043	0.081	0.37		
	0.069	10.9	0.186	21.6	0.141	13.3		

Note: For each model and component, a matrix $\left(\begin{smallmatrix}\alpha & \beta \\ \nu & \sigma\end{smallmatrix}\right)$ is given; α, β are the CES parameters, ν is the component's weight, and σ is the standard deviation of its member's choices in treatment 8. The remaining information for all models and all numbers of components are supplementary material.

significance of violating quantitative robustness in relation to random behavior, and in the case of ordered GEV to a p-value greater than the threshold .01.

Briefly, note that the relationship between random utility and random behavior is inverted for the plain models with standard error specification (structural regression and multinomial logit). As for structural regression, restricting errors to be normal improves qualitative robustness (five components out-of-sample instead of four) and avoids significant violations of quantitative robustness for all folds. This corresponds with the previous discussion, as restricting errors to be normal decreases the degrees of freedom in fitting purely statistical patterns, which increases robustness, albeit robustness at a comparably weak goodness-of-fit overall. As for multinomial logit, the restriction to extreme value perturbations restricts the goodnes-of-fit in all dimensions, including predictive accuracy. This suggests that the generalized extreme value perturbations in ordered GEV do not simply fit statistical patterns but allow for closer association of the leptokurtic choice patterns with the underlying utility differences.

As for random taste models, the reason for the lacking robustness is a little different. By construction, the model relates the choice pattern to utility differences, albeit to differences in utility parameters rather than utilities itself. Hence, it adapts to variations of opportunity costs and violates parametric robustness in fold S_1 less significantly than random behavior. Random taste models do not allow to fit "weak Leontief" subjects, however, i.e. component 4 for ordered GEV in Table 5. Weak Leontief subjects choose noisily around the minimax choice, with $\beta < -50$ in their utility functions $\big((1-\alpha)\pi_i^\beta + \alpha\pi_j^\beta\big)^{1/\beta}$. If $\beta < -50$, however, the weights α are irrelevant,[17] and hence the noisiness of weak Leontief subjects cannot be explained by variations of α. In this sense, random taste models are behaviorally incomplete and their approximation of weak Leontief subjects is not robust out-of-sample. As Table 5 shows, these weak Leontief subjects are approximated in-sample by component number 6, which is indistinct from component 5 (weak Cobb-Douglas) out-of-sample. This failure to fit weak Leontief choices is quantitatively significant when the random taste model has to predict behavior in folds S_1 and S_2 (see Table 4c), where the Leontief choices are more pronounced than in the other treatments.

The observation that random behavior and random taste models fail to accurately fit parts of the choice pattern does not imply that these models overfit significantly in relation what is expected by the BIC correction. The degree of overfitting is the difference $LL_{\text{In}} - LL_{\text{Out}}$, and I consider it significant if it exceeds the BIC correction term significantly (Vuong, 1989, Eq. 5.9). Using this criterion, the amount of overfitting is insignificant for all structural models (Table 4c).

Result 6.4 (Overfitting)**.** *Not one of the structural models overfits significantly (*$p < .01$*) if the number of components is chosen adequately (*$K \geq 4$*).*

This confirms that the structural models do not overfit in excess of the amount that is to be expected by adding further parameters. In random behavior and random taste models, the added components tend to become imprecise and indistinct out-of-sample, but they are not systematically wrong. This differs in the case of atheoretic regression, as discussed shortly, and confirms the conjecture voiced in the literature (Keane, 2010a; Rust, 2010) that structural modeling largely prevents overfitting—very much in contrast to atheoretic modeling.

The required dimensionality $K \geq 4$ is predicated by *ICL-BIC* in-sample and thus

[17] Note that $\big((1-\alpha)\pi_i^\beta + \alpha\pi_j^\beta\big)^{1/\beta}$ converges to $\min\{\pi_i, \pi_j\}$ regardless of α when β tends to $-\infty$.

known without cross validation. In addition it is intuitive. It allows for the two components with high-precision utility maximizers, egoistic and Leontief, for one component with low-precision members (component 3 in the random behavior model has standard deviation of 35, see Table 5), and for one component with intermediate precision (type 4 in the random behavior model has standard deviation 15 out of sample). Nonetheless, the observation that overfitting is insignificant for *all* structural models with at least this dimensionality deserves emphasis. It suggests that six observations per subject suffice to train the structural models, although the parameter space in the analyzed structural models has a dimensionality that would be considered high by most standards (up to 35 parameters for 176×8 observations).

Finally, the following result summarizes the main observations on the two atheoretic regression models.

Result 6.5 (Atheoretic regression)**.** *The linear regression models overfit highly significantly and improvements of the in-sample fit—achieved by adding either interaction terms or mixture components—are not robust to treatment-based cross validation.*

Apparently, the descriptive accuracy achieved by linear (atheoretic) regression is misleading in the analyzed dictator games. This again confirms the conjecture voiced in the literature. The reason is rather obvious: The non-linearity of the best-response functions in most treatment parameters (in all parameters but endowment E, that is) cannot be approximated by linear functions and interactions. Since most subjects tend to maximize utility indeed, as shown by Andreoni and Miller (2002), their precision is too high to be approximated linearly. While the choice patterns seem to be captured reasonably well in-sample, although worse than with the advanced structural models, the predictions are systematically off the mark out-of-sample. In particular, regression models fail to predict the location of the second mode in the empirical distributions, i.e. the choices made by the high-precision Leontief subjects. These systematic mistakes, in turn, imply the significance of overfitting. Having said that, it seems fair to preview that regression works well in public goods games.

7 Modeling choice in public goods games

Analyses of public goods games based on random utility models are fairly common in the experimental literature, but always under the premise that IIA is satisfied.[18] Thus, it will be interesting to see how relevant the above observation that IIA need be relaxed to capture excess kurtosis proves to be in this context. Random behavior modeling was used by Bardsley and Moffatt (2007), and only they account for subject heterogeneity by finite mixture modeling. In turn, random taste models or GEV models are not considered in the existing literature.

The following analysis comprises mixture models with up to four components, as the out-of-sample fit LL_{Out} drops beyond three components for all but the random utility models, and *ICL-BIC* drops beyond four components for the random utility models (details on this are provided in the supplementary material). The basic (atheoretic) regression model regresses contributions on treatment parameters again (group size, external return, and internal return), while the extended regression model additionally includes interaction terms. Table 6 summarizes the in-sample fit for all models and Table 7 describes their out-of-sample fit.[19] The main results on the in-sample fit are summarized first.

Result 7.1 (Descriptive accuracy)**.** *There are three tiers of models again. First, regression, random behavior, and ordered GEV fit about similarly well (none of them fits significantly worse than any other of them at* $\alpha = .01$*). Second, multinomial logit fits significantly worse than ordered GEV, and lastly, random taste does not fit.*

Similarly to modeling dictator game choices, the ability to adjust kurtosis in relation to multinomial logit proves important, as implied by the significance of the difference between multinomial logit and ordered GEV. Contrary to above, the fitted distributions tend to be platykurtic rather than leptokurtic for about 70% of the subjects, i.e. they exhibit flatter peaks and thinner tails. This corresponds loosely with $\rho = 10$ in Figure 2. Ordered GEV models with $\rho > 1$ do not satisfy the sufficient condition for being random utility models (McFadden, 1978), however. Intuitively, the

[18] For example, multinomial logit was applied by Anderson et al. (1998) to standard public goods games, by Offerman et al. (1998), Myatt and Wallace (2008), and Choi et al. (2008) to threshold public goods games, and by Willinger and Ziegelmeyer (2001) and Yi (2003) to nonlinear games.

[19] Again, the supplementary material contains all parameter estimates and plots showing that the empirical distributions are fitted well also qualitatively.

Table 6: In-sample summary for public goods games

(a) Goodness of fit $\mathit{ICL\text{-}BIC} = -LL + d/2 \cdot \ln O + \mathrm{En}(\hat{\tau})$ (less is better)

		Number of components			
	#Pars	One	Two	Three	Four
Regression	5	996	903	881	**876**
Extended regression	8	1000	911	894	**893**
Structural regression	3	995	905	**891**	897
Random behavior	4	996	908	**895**	900
Random taste	4	1175	1170	**1098**	1107
Random utility	3	1018	935	941	**916**
Ordered GEV	4	1005	922	919	**893**

(b) Likelihood-ratio tests using the in-sample-optimal number of components

		Model 2						
Model 1	CLC	Regr 4	ExtReg 4	StructReg 3	RBehav 3	RTaste 3	RUtil 4	OGEV 4
Regr 4	835.67		=	=	=	>>>	>	=
ExtReg 4	832.2	=		=	=	>>>	=	=
StructReg 3	872.23	=	=		=	>>>	=	=
RBehav 3	871.03	=	=	=		>>>	=	=
RTaste 3	1073.97	<<<	<<<	<<<	<<<		<<<	<<<
RUtil 4	876.04	<	=	=	=	>>>		<<<
OGEV 4	852.74	=	=	=	=	>>>	>>>	
Baseline 1	1043	<<<	<<<	<<<	<<<	>>>	<<<	<<<

$H_0 : LL(\mathit{Model\ 1}) = LL(\mathit{Model\ 2})$ with $>>>, >>, >$ indicating $p \leq .001, .01, 0.1$ in favor of Model 1. For a detailed description, see Table 3b.

flatness of the peaks implies that deviations from the utility maximizer are comparatively likely, i.e. precision is low. The probability of remote choices has to be proportional, though, while the fitted choice probabilities actually drop supra-proportionally outside the neighborhood of the utility maximizer.

From a more general point of view, platykurtic choice patterns suggest that subjects have an idea of their utility maximizer and randomize fairly uniformly over its neighborhood. The corresponding two thirds of the subjects (with $\rho \approx 10$) are therefore closer to random behavior than random utility. The more relevant observation is, however, that ordered GEV offers the flexibility required to capture such choice patterns (by allowing for $\rho > 1$). Thus, although it does not improve upon the explicit random behavior models with respect to goodness-of-fit here, it provides a general framework to model social donations. In contrast, the basic random utility model (multinomial logit) and the random taste model cannot capture these behavioral patterns. In particular, rationalizing choices by adjusting the altruism coefficient α in cases where the utility maximizer is not sufficiently salient proves econometrically ineffective.[20]

Another difference to dictator games is that the linear approximation used in regression models seems adequate here. Arguably, the reason is the previous observation that the utility maximizer is not salient, which implies that noise is comparably high and that choice patterns can be approximated linearly without much loss. To quantify the noisiness of choices, let us look at the hypothetical model that predicts the actually observed relative frequencies in all treatments. It would score a log-likelihood of -787 and a pseudo-R^2 of 0.245. The corresponding pseudo-R^2 in the dictator game data is 0.579. That is, the upper bar for the descriptive accuracy is rather low to begin with. In relation to this upper benchmark of -787, in turn, the in-sample likelihood -850 of ordered GEV is good. In order to ultimately assess the fit of atheoretic regression, let us look at the out-of-sample results next (Table 7).

Result 7.2 (Predictive accuracy). *All three modeling approaches (atheoretic regression, random behavior, and random utility) are in principle predictive. However, multinomial logit predicts significantly worse than ordered GEV, and the extended regression model (which includes interaction terms) predicts significantly worse than the plain regression model. None of these models overfits systematically (at $\alpha = .01$) if the number of components is chosen appropriately (i.e. to minimize ICL-BIC).*

Thus, in contrast to dictator games, linear approximations suffice in-sample and do not overfit more than predicted by the BIC correction term. The number of parameters of the extended model is rather large, however, and thus its BIC correction is large. Indeed, the model without interaction terms fits significantly better than the extended model out-of-sample, and in this sense, the extended model does overfit. Further, in absolute terms, the regression models overfit slightly more than the structural models, i.e. random behavior and ordered GEV.

Finally, let us look at the robustness of the identified components again.

[20] To be more precise, strict rationalization of all choices by varying altruism is ineffective, because the general amount of noise is too high. To see this, compute the $\alpha' = \alpha/(1-\alpha)$ that explain the mean observations in the various treatments. The ratio of the highest α' to the lowest α' required is 3.08 in dictator games, and 9.05 in public goods games. Thus, to explain the public goods data, a relatively large variance of α is required at the individual level, and this in turn overshoots in the treatments where the external return is very high.

Table 7: Summary on the out-of-sample analysis of public goods games

(a) The out-of-sample log-likelihoods Eq. (14) and optimal number of components

	Number of components						
Model 1	1		2		3		4
Regr	−996.75	$\ll$	**−950.35**	=	−941.04	=	−944.05
ExtReg	−1004.76	<	**−969.41**	=	−963.48	=	−974.75
StructReg	−993.13	<	−962.68	<	**−950.3**	=	−959.03
RBehav	−994.23	$\ll$	**−948.79**	=	−943.28	=	−939.4
RTaste	−1291.26	=	−1304.69	$\lll$	**−1184.16**	=	−1184.47
RUtil	−1016.17	=	−1019.12	<	−1014.07	<	**−983.43**
OGEV	−1003.39	=	−1041.86	$\ll$	−985.15	<	**−945.39**

Note: $\lll, \ll, <$ indicate rejection of $H_0 : LL_{\text{Out}}(k\ \textit{Components}) = LL_{\text{Out}}(k+1\ \textit{Components})$ with $p <$ $.001, .01, .1$. The log-likelihood LL_{Out} for the *out-of-sample optimal* number of components (smallest number of components that is not significantly improved upon at $\alpha = 0.1$) is set in bold-face type.

(b) Likelihood-ratio tests of LL_{Out} using the out-of-sample-optimal number of components

		Model 2						
Model 1	LL	Regr 2	ExtReg 2	StructReg 3	RBehav 2	RTaste 3	RUtil 4	OGEV 4
Regr 2	−950.35		>>>	=	=	>>>	>	=
ExtReg 2	−969.41	<<<		=	<	>>>	=	=
StructReg 3	−950.3	=	=		=	>>>	>	=
RBehav 2	−948.79	=	>	=		>>>	>>	=
RTaste 3	−1184.16	<<<	<<<	<<<	<<<		<<<	<<<
RUtil 4	−983.43	<	=	<	<<	>>>		<<<
OGEV 4	−945.39	=	=	=	=	>>>	>>>	
Baseline 1	−1043	<<	<	<	<<	>	=	<

$H_0 : LL_{\text{Out}}(M1) = LL_{\text{Out}}(M2)$ with $>>>, >>, >$ indicating $p \leq .001, .01, .1$ in favor of M1. For a detailed description, see Table 4b.

(c) Likelihood-ratio tests on overfitting and robustness

	LR tests on overfitting					LR tests on robustness for folds			
Model 1	1	2	3	4		S_1	S_2	S_3	S_4
Regr	=	=	=	=	Regr 4	=	>	>	=
ExtReg	=	>	=	=	ExtReg 4	>	>>>	>	>
StructReg	=	>>	=	=	StructReg 3	>	=	=	>
RBehav	=	=	=	=	RBehav 3	=	=	=	>
RTaste	>>>	>>>	>>	>>>	RTaste 3	>>>	>	=	=
RUtil	=	>>	>	>	RUtil 4	=	>>	=	>
OGEV	=	>>	>	=	OGEV 4	=	>	=	=

$H_0 : LL_{\text{In}}(\textit{Model 1}) = LL_{\text{Out}}(\textit{Model 1})$, see also Table 4c.

Result 7.3 (Qualitative robustness)**.** *Both random utility models (multinomial logit and ordered GEV) achieve precise in-sample and robust out-of-sample distinction of components. In all other cases, the identification is either less precise in-sample ($K < 4$) or not robust to cross validation.*

Although the ordered GEV model adds neither descriptive nor predictive accuracy in the noisy environment of public goods games, it provides robustness of identified motives. Again, qualitative robustness applies for both kinds of random utility models considered here, and in both dictator games and public goods games, but not for any other model in either of these games. Thus, qualitative robustness of identified motives is an attribute of random utility approaches, i.e. of relating choice patterns to utility differences, rather than random behavior models, which match their statistical properties. As for quantitative robustness (Table 7c), allowing for generalized extreme value perturbations is more adequate than multinomial logit again, which corroborates the above discussion.

The robustness of ordered GEV seems to come at a price, as the random utility models catch up with the statistically more flexible random behavior models in terms of the in-sample fit (*ICL-BIC*) only as the fourth component is introduced. Intuitively, this may be inevitable if we seek robustness of identified components, but conversely, it may also be taken as a suggestion for further research. This is briefly discussed toward the end of the next section.

8 Conclusion

Recent experimental studies provide quantitative estimates of utility parameters (for example Cappelen et al., 2007; Cox et al., 2007; Fisman et al., 2007). In contrast to earlier studies, these studies estimate utility in relation to models of choice, in order to obtain joint estimates of choice and utility functions, which interact in individual decision making (Andreoni and Miller, 2002; Goeree et al., 2002). Their estimates differ, however, as they use different models of choice, and in general, utility estimates differ across studies and even across treatments (Blanco et al., 2011). Such disarray may possibly be avoided if parameters are estimated in relation to a "homeomorphic" model of choice, i.e. a model that exhibits both maximal descriptive fit and insignificant overfitting. In contrast to choice under risk (Wakker, 2010), however, the econometric structure of choice is not well understood in experimental games.

The present study addresses this void by analyzing the structure of choice in two widely used experimental games, dictator games and public goods games. These are both games of social donations, and with respect to the precision in utility maximization, their combination covers a wide range of choice patterns. Dictator games elicit rather precise and leptokurtic choice patterns, while public goods games elicit more uniform and platykurtic choice patterns. In typical applications, it may be not clear (a priori) where in this range a given game is located, and hence, the identification of a generally valid model of choice seems desirable. The ordered GEV model of Small (1987), a random utility model for ordered alternatives, comes fairly close to be such a model.

The general results are that structural models avoid overfitting (in the "simple" games analyzed here), that structural models with generalized errors (to adjust kurtosis) generally fit well, and that random utility models yield robust identification of components (social motives). These qualitative observations match with ordered GEV, which also fits best quantitatively (i.e. uniquely best in some dimensions and jointly so in other dimensions). As random utility models are used frequently and the ordering of alternatives is natural, ordered GEV is also an appealing model in this context, but surprisingly it is yet to be applied in experimental studies.

In contrast, the multinomial logit model of random utility generally fits worse because it cannot adjust to excess kurtosis in high-precision dictator games or to platykurtosis in low-precision public goods games, random behavior and random taste models do not identify components robustly, and (linear) regression models do not fit in high-precision games. The results suggest that modeling donations and contributions by ordered GEV will help to alleviate the current issues in utility estimation for games comparable to dictator and public goods games, and that similarly robust models may exist in alternative circumstances.

As future research, there seems to be room for improvement in the design of choice models and for extending the scope of the analysis. As for modeling choice, ordered GEV has only limited capacity to adjust to platykurtic choice patterns. In contrast to normally distributed random behavior, the peaks are comparably thin even for $\rho \approx 10$ (see Figure 2), which seems to limit its capability to describe low-precision subjects. Thus, it is conceivable that alternative utility-based models can be developed that improve upon ordered GEV.

As for the scope of the analysis, it may be extended by investigating choice patterns (i.e. predictive accuracy, kurtosis, and robustness of identified components) in alternative circumstances. Such extensions will probably require joint modeling of the beliefs underlying strategic choice, e.g. by quantal response equilibria (McKelvey and Palfrey, 1995) such as in the present analysis, level-*k* modeling (Stahl and Wilson, 1995), or noisy introspection (Goeree and Holt, 2004). For recent comparative analyses, see e.g. Crawford and Iriberri (2007a,b) and Breitmoser (2012). Additionally, one may extend the analysis by analyzing robustness across experiments (i.e. across subject pools). The ordered GEV model, which proved robust across treatments, may be a natural starting point, but additional issues such as the quantification of differences between subject pools will have to be resolved. Potentially, when strategic choice is understood in a larger variety of games, cross-comparisons across experimental analyses will be reliable.

References

Amemiya, T. (1978). On a two-step estimation of a multivariate logit model. *Journal of Econometrics*, 8(1):13–21.

Anderson, S. and De Palma, A. (1992). Multiproduct firms: a nested logit approach. *The Journal of Industrial Economics*, pages 261–276.

Anderson, S., Goeree, J., and Holt, C. (1998). A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics*, 70(2):297–323.

Anderson, S., Goeree, J., and Holt, C. (2001). Minimum-effort coordination games: Stochastic potential and logit equilibrium. *Games and Economic Behavior*, 34(2):177–199.

Anderson, S., Goeree, J., and Holt, C. (2002). The logit equilibrium: A perspective on intuitive behavioral anomalies. *Southern Economic Journal*, 69(1):21–47.

Andreoni, J. (1995a). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, 85(4):891–904.

Andreoni, J. (1995b). Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, 110(1):1–21.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Andreoni, J. and Scholz, J. (1998). An econometric analysis of charitable giving with interdependent preferences. *Economic Inquiry*, 36(3):410–428.

Arcidiacono, P. and Jones, J. (2003). Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica*, 71(3):933–946.

Auger, A., Hansen, N., Perez Zerpa, J., Ros, R., and Schoenauer, M. (2009). Experimental comparisons of derivative free optimization algorithms. *Experimental Algorithms*, pages 3–15.

Auten, G., Sieg, H., and Clotfelter, C. (2002). Charitable giving, income, and taxes: an analysis of panel data. *American Economic Review*, 92(1):371–382.

Bakija, J. and Heim, B. (2011). How does charitable giving respond to incentives and income? new estimates from panel data. *National Tax Journal*, 64(2):615–50.

Bardsley, N. and Moffatt, P. (2007). The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, 62(2):161–193.

Battalio, R., Samuelson, L., and Huyck, J. (2001). Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica*, 69(3):749–764.

Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Blanco, M., Engelmann, D., and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

Breitmoser, Y. (2012). Strategic reasoning in p-beauty contests. *Games and Economic Behavior (in press).*

Bruhin, A., Fehr-Duda, H., and Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78(4):1375–1412.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Princeton University Press.

Cappelen, A., Hole, A., Sørensen, E., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A., Hole, A., Sørensen, E., and Tungodden, B. (2010a). Modeling individual choices in experiments: Random utility versus random behavior. *Working paper.*

Cappelen, A., Moene, K., Sørensen, E., and Tungodden, B. (2011). Needs vs entitlements: an international fairness experiment. *Journal of the European Economic Association.*

Cappelen, A., Sørensen, E., and Tungodden, B. (2010b). Responsibility for what? Fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Card, D., DellaVigna, S., and Malmendier, U. (2011). The role of theory in field experiments. *The Journal of Economic Perspectives*, 25(3):39–62.

Choi, S., Gale, D., and Kariv, S. (2008). Sequential equilibrium in monotone games: A theory-based analysis of experimental data. *Journal of Economic Theory*, 143(1):302–330.

Conte, A., Hey, J., and Moffatt, P. (2011). Mixture models of choice under risk. *Journal of Econometrics*, 162(1):79–88.

Conte, A. and Moffatt, P. (2009). The pluralism of fairness ideals: a comment. *Working paper.*

Conte, A. and Moffatt, P. (2010). The econometric modelling of social preferences. *Jena Economic Research Papers No. 2010-042*.

Costa-Gomes, M., Crawford, V., and Iriberri, N. (2009). Comparing models of strategic thinking in Van Huyck, Battalio, and Beil's coordination games. *Journal of the European Economic Association*, 7(2-3):365–376.

Cox, J., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1):17–45.

Crawford, V. and Iriberri, N. (2007a). Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" games. *American Economic Review*, 97(5):1731–1750.

Crawford, V. and Iriberri, N. (2007b). Level-*k* auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *The American Economic Review*, 97(5):1858–1876.

Forsythe, R., Horowitz, J., Savin, N., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Goeree, J. and Holt, C. (2004). A model of noisy introspection. *Games and Economic Behavior*, 46(2):365–382.

Goeree, J., Holt, C., and Laury, S. (2002). Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, 83(2):255–276.

Hansen, N. and Kern, S. (2004). Evaluating the cma evolution strategy on multimodal test functions. *Parallel Problem Solving from NaturePPSN VIII*, 3242(x):282–291.

Hansen, N., Müller, S., and Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary Computation*, 11(1):1–18.

Harrison, G. and Rutström, E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*, 12(2):133–158.

Hey, J. (2005). Why we should not be silent about noise. *Experimental Economics*, 8(4):325–345.

Hey, J., Lotito, G., and Maffioletti, A. (2010). The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of risk and uncertainty*, 41(2):81–111.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3):346–380.

Keane, M. (2010a). A structural perspective on the experimentalist school. *The Journal of Economic Perspectives*, 24(2):47–58.

Keane, M. and Wolpin, K. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior. *International Economic Review*, 48(4):1351–1378.

Keane, M. P. (2010b). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1):3–20.

Keser, C. and van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1):23–39.

Kübler, D. and Weizsäcker, G. (2004). Limited depth of reasoning and failure of cascade formation in the laboratory. *Review of Economic Studies*, 71(2):425–441.

Kübler, D. and Weizsäcker, G. (2005). Are longer cascades more stable? *Journal of the European Economic Association*, 3(2-3):330–339.

Lee, B. (1999). Calling patterns and usage of residential toll service under self selecting tariffs. *Journal of Regulatory economics*, 16(1):45–82.

Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.

Loomes, G. (2005). Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. *Experimental Economics*, 8(4):301–323.

Luce, R. (1959). *Individual choice behavior*. Wiley New York.

McFadden, D. (1973). Conditional logit analysis of qualitative choice models. *Frontiers of Econometrics, ed. P. Zarembka. New York: Academic Press*, pages 105–142.

McFadden, D. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement*, 5(4):363–390.

McFadden, D. (1978). Modelling the choice of residential location. In Karlqvist, A., Lundqvist, L., Snickars, F., and Weibull, J., editors, *Spatial interaction theory and planning models*, pages 75–96. North Holland, Amsterdam.

McFadden, D. (1981). Econometric models of probabilistic choice. In Manski, C. and McFadden, D., editors, *Structural analysis of discrete data with econometric applications*, pages 198–274. MIT Press, Cambridge.

McKelvey, R. and Palfrey, T. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in probability and statistics.

Moré, J. and Wild, S. (2009). Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191.

Myatt, D. and Wallace, C. (2008). An evolutionary analysis of the volunteer's dilemma. *Games and Economic Behavior*, 62(1):67–76.

Nevo, A. and Whinston, M. (2010). Taking the dogma out of econometrics: Structural modeling and credible inference. *The Journal of Economic Perspectives*, 24(2):69–81.

Offerman, T., Schram, A., and Sonnemans, J. (1998). Quantal response models in step-level public good games. *European Journal of Political Economy*, 14(1):89–100.

Payne, A. (1998). Does the government crowd-out private donations? new evidence from a sample of non-profit firms. *Journal of Public Economics*, 69(3):323–345.

Powell, M. (2008). Developments of newuoa for minimization without derivatives. *IMA journal of numerical analysis*, 28(4):649.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302.

Rosenthal, R. (1989). A bounded-rationality approach to the study of noncooperative games. *International Journal of Game Theory*, 18(3):273–292.

Rust, J. (2010). Comments on: "structural vs. atheoretic approaches to econometrics" by Michael Keane. *Journal of Econometrics*, 156(1):21–24.

Samuelson, L. (1985). On the independence from irrelevant alternatives in probabilistic choice models. *Journal of Economic Theory*, 35(2):376–389.

Schokkaert, E. (2006). The empirical analysis of transfer motives. *Handbook on the Economics of Giving, Reciprocity and Altruism*, 1:127–181.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Small, K. (1987). A discrete choice model for ordered alternatives. *Econometrica*, 55(2):409–424.

Small, K. (1994). Approximate generalized extreme value models of discrete choice. *Journal of Econometrics*, 62(2):351–382.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72.

Stahl, D. and Wilson, P. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge Univ Pr.

Train, K., McFadden, D., and Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *The Rand Journal of Economics*, pages 109–123.

Turocy, T. (2005). A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51(2):243–263.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281–299.

Vovsha, P. (1997). Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. *Transportation Research Record*, 1607(-1):6–15.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wakker, P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge Univ Pr.

Whitten, G. and Palmer, H. (1996). Heightening comparativists' concern for model choice: Voting behavior in great britain and the netherlands. *American Journal of Political Science*, 40(1):231–260.

Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. *Risk aversion in experiments*, 12:197–292.

Wilcox, N. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, 162(1):89–104.

Willinger, M. and Ziegelmeyer, A. (2001). Strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. *Experimental Economics*, 4(2):131–144.

Yi, K. (2003). A quantal response equilibrium model of order-statistic games. *Journal of Economic Behavior and Organization*, 51(3):413–425.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21(1):299–313.