
Munich Personal RePEc Archive

Structural modeling of altruistic giving

Breitmoser, Yves

EUV Frankfurt (Oder)

5 August 2010

Online at https://mpra.ub.uni-muenchen.de/24262/
MPRA Paper No. 24262, posted 06 Aug 2010 10:05 UTC

Structural modeling of altruistic giving

Yves Breitmoser*
EUV Frankfurt (Oder)

August 5, 2010

Abstract

The paper analyzes econometric models of altruistic giving in dictator and public goods games. Using existing data sets, I evaluate internal and external validity of "atheoretic" regression models as well as structural models of random behavior, random coefficients, and random utility, controlling for subject heterogeneity by finite mixture modeling. In dictator games, atheoretic regression lacks external validity, while random coefficient models and random utility models offer high degrees of both internal and external validity. In public goods games, regression works comparably well, being bettered only by random utility models. Overall, the ordered GEV model of random utility is most appropriate to describe choices in the considered games.

JEL–Codes: C44, C50, C72, D64
Keywords: structural modeling, altruism, dictator game, public goods, ordered choice sets

*I thank Friedel Bolle for his helpful comments and James Andreoni and John Miller for permitting me to use 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

Numerous studies have investigated how the actions of experimental subjects relate to game theoretic predictions. It was found that experimental subjects generally deviate from Nash equilibrium, which raised the questions whether the deviations are systematic and how they could be explained. Several strands of literature emerged (for a more complete survey of this research, see Camerer, 2003). One of them investigated whether deviations from the predictions can be the result of social preferences (e.g. Rabin, 1993, and Levine, 1998), another one additionally allowed that subjects play noisy responses in relation to some utility function (e.g. Rosenthal, 1989, and McKelvey and Palfrey, 1995), a third strand investigated how choice patterns depended on circumstances (Forsythe et al., 1994; Hoffman et al., 1994; Andreoni, 1995b), and another one investigated the consistency of individual choices (e.g. Andreoni, 1995a). It was found that both social preferences and noisy responses seem to explain the observed deviations from equilibrium predictions, in response to which Andreoni and Miller (2002) and Goeree et al. (2002) conducted experiments to separate these potential explanations. They systematically varied exchange rates in dictator games and public goods games (respectively) to get a complete overview of individual choice patterns. The conclusions were that social preferences and noisiness of responses individually interact.

This, in turn, led researches to estimate structural models of altruistic giving that contain both social preferences and a source of randomness inducing noisy responses. Fisman et al. (2007) assumed that subjects deviate stochastically from their individual best response (*random behavior*), Cox et al. (2007) assumed that the altruism coefficient in the individual utility function is fluctuating randomly (*random coefficient*), and Cappelen et al. (2007) considered a multinomial logit model of choice (*random utility*). This multitude of approaches has an obvious flaw, as Conte and Moffatt (2009) showed that the estimated motive of giving depends on the model chosen (they do so by fitting a random behavior model to Cappelen et al.'s data).[1]

In relation to this literature, the present paper answers two questions: Which of

[1] Conte and Moffatt also criticize the random utility model chosen by Cappelen et al. for neglecting the orderedness of the choice set in dictator games.

the three approaches toward structural modeling of altruistic giving is most valid? And, do the internal or external validity gained justify or even necessitate the move from atheoretic regression toward structural modeling? To this end, I revisit the experimental data of Andreoni and Miller (2002) and Goeree et al. (2002), estimate the models to be discussed, and compare measures of internal validity (BIC in-sample) and external validity (LL out-of-sample). The main results are that random utility modeling is in general most valid, random coefficient models may fit well (as they do in Andreoni and Miller's dictator games) but they may also fail drastically (in Goeree et al.'s public goods games), and in general the move toward structural modeling is justified and may be necessary (the validity increases drastically in dictator games). The approach that emerges as most valid in our analysis is the ordered GEV model of random utility (Small, 1987), which has been overlooked by previous game-theoretic analyses, but poses a direct response to the critique that random utility as in multinomial logit ignores orderedness of choice sets (e.g. Conte and Moffatt, 2009).

From a more general point of view, the present paper contributes to the discussion of the comparative advantages of structural modeling and regression—recently revamped by e.g. Keane (2010) and Rust (2010)—by presenting quantitative evidence based on experimental data. As indicated, the evidence underlines the general idea that structural models have higher external validity than (linear) regression, but surprisingly they also have higher internal validity. That is, structural models (and in particular random utility models) are better in describing the basic characteristics of individual choice in the considered games, and hence they are preferable even if only internal validity is of interest.

The remainder is organized as follows. Section 2 introduces the dictator game data of Andreoni and Miller (2002) and discusses tobit regression models. Section 3 analyzes the standard structural models discussed in the literature, and Section 4 extends the analysis to random utility models relaxing the assumption of independence from irrelevant alternatives (i.e. ordered GEV and nested logit). Section 5 verifies the robustness of the dictator game results by analyzing the public goods game data of Goeree et al. (2002). Section 6 concludes. There is extensive supplementary material that lists (amongst others) all parameter estimates.

2 The data and initial analysis

In a dictator game, only player 1 has to choose a strategy. His choice affects the payoff of 2, however, and this payoff interdependence seems utility relevant for laboratory subjects. Let 1's strategy set be denoted as $S_1 = \{0, 1, \ldots, B\}$, with B as endowment, and let τ_1, τ_2 denote (positive) exchange rates. The two players' payoffs are

$$\pi_1(s) = \tau_1(B - s) \qquad \pi_2(s_1) = \tau_2 s \qquad \forall s \in S_1. \tag{1}$$

We analyze the dictator game experiment conducted by Andreoni and Miller (2002), who explicitly designed their experiment to evaluate the consistency of dictator decisions with utility maximization. There are eight decisions per subject, based on systematic variations of (B, τ_1, τ_2). This allows us to econometrically disentangle randomization and distributive preferences at the individual level.

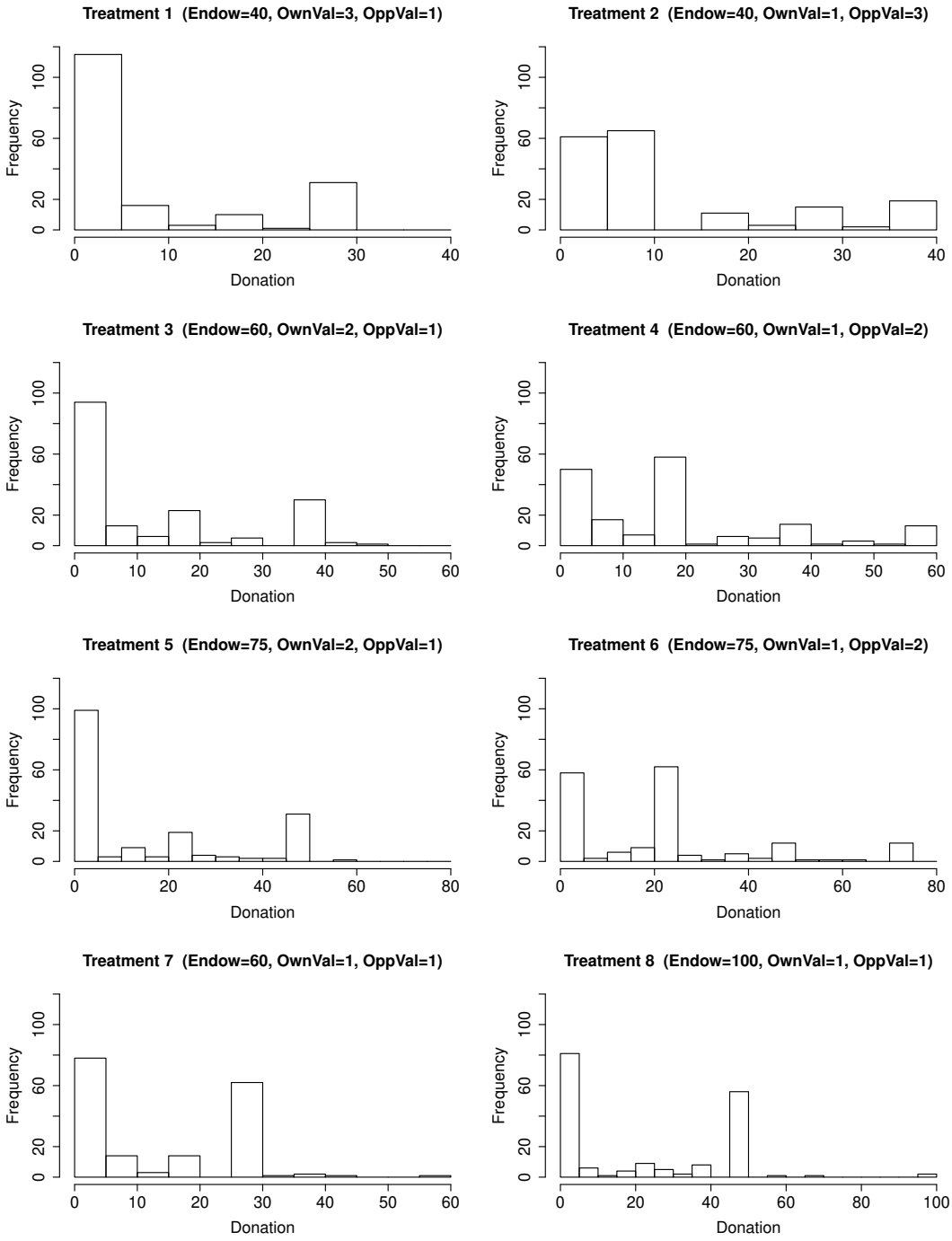
Figure 1 provides an overview of the treatment parameters and the data. The data exhibits typical characteristics of dictator games. For example, the donations are fairly moderate overall, they are decreasing in τ_1, and they are increasing in τ_2. The standard analytical approach to show this is to estimate a (tobit) regression model of donations on (B, τ_1, τ_2). In our context, the result is

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

where ε is normal with standard deviation $\hat{\sigma} = 27.645$. Regression analyses of this kind are common in experimental studies, arguably because they are pure—by avoiding complex specification—which seems to imply that the data speaks for itself. In our case, the analysis suggests that the aforementioned effects are significant (s_1 is decreasing in τ_1 and increasing in τ_2), and that the donations fall by 4.3 on average per unit increase of τ_1 and they increase by 4.9 on average per unit increase of τ_2. The validity of such conclusions is questionable, however, as tobit models ignore the structure of dictator games. Hence, their results are susceptible to what originally became known as the Lucas critique. This issue is the topic of the following.

Throughout, we distinguish external validity and internal validity. Results have internal validity if they accurately reflect the observed interaction, and they have external validity if they continue to hold in related circumstances. Arguably, regression

Figure 1: Treatment parameters and data of Andreoni and Miller (2002)



models have higher internal validity than structural models, whereas structural models have higher external validity. The former may result, since regression models are estimated without restrictive structural assumptions to be obeyed, while the latter follows precisely from the structural restrictions, which reduce the risk of overfitting and improve one's chances of capturing the structure of decision making. This depends, of course, on how well one's model structure approximates the problem structure.

In the present context, it seems reasonable not to ignore subject heterogeneity. For example, Andreoni and Miller's analysis isolated seven types of subjects (as will be discussed in more detail below). In order to represent (latent) heterogeneity, we resort to finite mixture modeling (see e.g. Peel and MacLahlan, 2000).[2] That is, subject heterogeneity is described non-parametrically by distinguishing up to seven discrete types, rather than parametrically by assuming a continuous distribution of types. This approach is adopted, as Andreoni and Miller identified seven distinctive types rather than a continuum of types.

Formally, let K denote the set of subject types in the population, e.g. $K = \{1,2,3\}$ in a three-type population, let $(\mu_k)_{k\in K}$ denote the type shares, and for all $k \in K$, let P_k denote the parameter profile characterizing type k. Now, if $o_{j,t}$ denotes the tth observation of subject $j \in J$ in the data set and if $\sigma(o_{j,t}|P_k)$ is the probability that i chooses $o_{j,t}$ conditional on being of type k, the log-likelihood of $o = (o_{j,t})$ is

$$LL(o|P) = \sum_{j\in J} \ln \sum_{k\in K} \mu_k \prod_t \sigma(o_{j,t}|P_k). \tag{3}$$

Using the finite mixture approach, I have estimated models of subject heterogeneity where each type is described by the linear (tobit) model defined in Eq. (2). Table 1 lists the goodness-of-fit of the various models, distinguishing internal and external validity. The measure of internal validity is Bayes' information criterion (BIC, see Schwarz, 1978) of the model fitted to the whole data set. The measure of external validity is the log-likelihood (LL) of the respective model fit to a subset of the

[2] In experimental economics, finite mixture models are best known from analyses of strategic reasoning, starting with Stahl and Wilson (1995), but have recently been extended to choice under risk (Conte et al., 2008; Harrison and Rutström, 2009; Bruhin et al., 2010), giving in dictator games (Cappelen et al., 2007, 2010), and donations to public goods (Bardsley and Moffatt, 2007).

Table 1: Validity of the linear behavioral model Eq. (2) in dictator games

	Benchmarks		Number of types of the linear behavioral model						
	Lower	Upper	One	Two	Three	Four	Five	Six	Seven
Internal	5814	2449	4276	3783	3672	3491	3439	3409	3392
External	1574	643	1198	1091	1149	1252	1171	1215	1277

Note: Internal validity is BIC $= -LL + (\#Pars)/2 \cdot \ln(\#Obs)$ of the model fitted to the whole data set, external validity is $-LL$ in treatments $6, 8$ of the model fitted to the restricted sample (1–5,7).

data (treatments 1–5 and 7) and evaluated in the other two treatments (6 and 8).[3] In addition, Table 1 reports two benchmark measures that will be referred to frequently. These benchmarks are defined as follows.

Definition 2.1 (Benchmarks)**.** The *upper benchmark* is the absolute value of the log-likelihood obtained by a model that predicts the actually observed relative frequencies of all actions in all treatments. The *lower benchmark* is the absolute value of the log-likelihood of predicting uniform randomization in all treatments.

Note that the upper benchmark reported is the strict upper benchmark of models assuming independence of choices between treatments. If latent subject heterogeneity as reported by Andreoni and Miller (2002) exists indeed, such independence assumptions are invalid and the upper bound is not strict. It is intended as an indication of what to expect from a "good" model. The following result summarizes Table 1 (the respective parameter estimates are provided as supplementary material).

Result 2.2. *Tobit models lack validity. Finite mixtures of tobit models induce negative correlation between internal and external validity ($\hat{\rho} = -0.36$), and in relation to the benchmarks, the internally best model attains 72.0% internal validity*[4] *and 31.9% external validity.*

Note the particularly low external validity of all estimated tobit models. Dictator game results derived from tobit (or related linear) regression models do not continue

[3] The treatments chosen for the out-of-sample tests are intermediate in the sense that the donation in relation to endowment is intermediate. We thus investigate external validity with respect to related, non-extreme circumstances.

[4] $0.72 = (5814 - 3392)/(5814 - 2449)$

to hold even in closely related dictator games. Structural models of behavior may allow us to avoid this pitfall.

3 Randomness of behavior, coefficients, and utility

Andreoni and Miller distinguished six specific subject types and one residual type. The six specific types have Cobb-Douglas, Leontief, or linear utility functions and either high or medium consistency (i.e. accuracy) in maximizing utilities. On the one hand, the three utility functions are special cases of CES utilities,

$$u_i(\pi_i,\pi_j) = \Big((1-\alpha)\cdot(1+\pi_i)^{\beta} + \alpha\cdot(1+\pi_j)^{\beta}\Big)^{1/\beta}, \tag{4}$$

if (π_i,π_j) denotes the payoff profile in question and using $u_i = -(\text{abs}(\dots))^{1/\beta}$ in case the base is negative. CES utilities have also been assumed in the existing structural models discussed soon. The varying degrees of accuracy, on the other hand, will be represented by varying the scale of noise in models based on these utility functions.

Depending on where the noise is assumed to enter decision making, one may distinguish three classes of structural models for dictator games: random behavior, random coefficients, and random utility. In analyses of dictator games, random behavior has been studied by Fisman et al. (2007) and Conte and Moffatt (2009), random coefficients have been studied by Cox et al. (2007), and random utility by Cappelen et al. (2007). The comparative advantages of these models have not yet been analyzed, however. In this section we analyze the validity of the models as they have been discussed by these authors. Alternative models are considered below.

To provide formal definitions, let $u(s|\alpha,\beta)$ denote i's utility from donating $s \in S_1$, and define $\text{BR}(\alpha,\beta) \in \arg\max_{s\in S_1} u(s|\alpha,\beta)$ as the (generically unique) utility maximizing donation of a subject with parameters (α,β).

Definition 3.1 (Random behavior)**.** The choice of i is a random variable $S_i = \text{BR}(\alpha,\beta) + \varepsilon$, censored at 0 and B, where ε is normal with mean zero and standard deviation σ.

Definition 3.2 (Random coefficient)**.** The choice of i is a random variable $S_i = \text{BR}(\alpha,\beta)$, where α is such that $\alpha' := \alpha/(1-\alpha)$ has density $f(\alpha') = \rho \exp\left\{(|\alpha - m|/s)^{\rho}\right\} / 2s\Gamma(1/\rho)$.

Table 2: Internal and external validity of basic structural models

	Number of types						
	One	Two	Three	Four	Five	Six	Seven
Random behavior	4349	3889	3817	3824	3248	3261	3214
	1238	1119	1096	1091	1102	1088	1049
Random coefficient	4353	3861	3757	2964	2730	2668	2687
	1235	1100	1075	904	815	811	813
Random utility	4724	3917	3253	3238	3030	3021	2973
	1316	1112	971	973	937	924	919

Note: As in Table 1, the top row (per model) contains the BIC $= -LL + (\#Pars)/2 \cdot \ln(\#Obs)$ measure of internal validity and the bottom row contains the $-LL$ measure of external validity.

The assumption that $\alpha' := \alpha/(1-\alpha)$ has exponential power distribution with mean m, scale s, shape ρ (i.i.d. for each decision) follows Cox et al. (2007). The implied probability that i chooses an action $s_i' \leq s_i$ is $F(\alpha^*)$ where F is the cdf and α^* is chosen such that $u_i(s_i|\alpha^*) = u_i(s_i+1|\alpha^*)$. See Cox et al. (2007, Appendix B) for further illustrations on the computational procedure.

Definition 3.3 (Random utility)**.** Player i maximizes the utility $\tilde{u} = \lambda u(s|\alpha,\beta) + \varepsilon$ where $\lambda \geq 0$ and ε has extreme value distribution (i.i.d. for all options $s \in S_1$).

The assumption that ε be extreme value distributed implies the multinomial logit choice probabilities.

$$\forall s \in S_1 : \qquad \Pr(s) = e^{\lambda u(s|\alpha,\beta)} \Big/ \sum_{s' \in S_1} e^{\lambda u(s'|\alpha,\beta)}. \tag{5}$$

Table 2 summarizes internal and external validity for these three classes of models. The full list of parameter estimates is provided as supplementary material. Before the results are summarized, let me briefly comment on the optimization procedure. The main issues to be resolved are the non-concavity of the likelihood function in finite mixture models, which follows from the interchangeability of types, and the high degree of non-linearity of the log-likelihood in many structural models. I adopted a variety of maximization methods, including Nelder-Mead and gradient based ones,

and many different starting values to ensure global convergence. Furthermore, numerical accuracy tends to be an issue in the summation underlying Eq. (5), as the numbers to be summed up can become large. This issue had been resolved by appropriately adapting the internal representation of numbers (further details are available upon request). Finally, the well-known issues with two-step estimators (see e.g. Amemiya, 1978, and Arcidiacono and Jones, 2003) were avoided by maximizing the full-information likelihood jointly over all parameters.

Result 3.4. *All three structural models induce positive correlation between internal and external validity, and overall the most valid model is the random coefficient model. In relation to the benchmarks (Def. 2.1), it attains (up to) 93.5% internal validity and 82.0% external validity.*

Recall that the random coefficient model is based on four parameters per subject type, whereas the other two (structural) models are based on three parameters per type. Hence, the improved goodness-of-fit measures of the random coefficient models may be due to their higher flexibility in fitting behavioral patterns. To address this possibility, we will next investigate slightly more flexible random utility models.

4 Generalized random utility models

Random utility models as they are applied in experimental analyses generally assume i.i.d. random components ε (as in Def. 3.3). This induces independence from irrelevant alternatives (IIA) in the choice probabilities, and as such it is an implication that is not generally realistic. The established choice theoretic approaches toward modeling deviations from IIA assume that subjects group choices with similar characteristics and that they first pick a group ("nest") and second pick a choice from that nest. Depending on whether the assumed nests overlap, we distinguish nested logit models and cross-nested logit models, which both are special cases of the case that the random components ε in Def. 3.3 have generalized extreme value (GEV) distribution (McFadden, 1978).

For some reason, GEV models of (strategic) choice have not yet been adopted in experimental economics. To my knowledge, all random utility analyses of exper-

imental data that followed McKelvey and Palfrey (1995) and Anderson et al. (1998) assume multinomial strategy sets. One reason may be that GEV models are computationally more intensive that multinomial logit (see e.g. Small, 1994), and a second reason may be that the additional flexibility attained in GEV models is feared to facilitate overfitting. The latter loosely relates to the result of Haile et al. (2008) who showed that random utility models may fit any data set if the distributional assumptions on ε are sufficiently weak. To be sure, the assumptions in standard GEV models are far more restrictive than the technical requirements of Haile et al., but empirical analyses similar to the one reported next seem necessary to convince practitioners.

The main issue in defining suitable GEV models is to identify characteristics based on which subjects group choices. Small (1987) argues that subjects nest choices based on proximity under the ordering of the choice set (if such an ordering exists) and defines the *ordered GEV* model to capture this possibility. Ordered GEV is a special case of cross-nested logit (see e.g. Vovsha, 1997) and of "elimination by aspect" (Tversky, 1972, see also McFadden, 1981, p. 225f). Alternatively, we also consider two disjointly nested models. The first one is a "control model" to verify whether seemingly arbitrary nesting based on numeral digits induces a good fit. Here, two choice options belong to the same nest if their first digits coincide.

$$s \text{ and } s' \text{ are in the same nest} \quad \Leftrightarrow \quad \lfloor s/10 \rfloor = \lfloor s'/10 \rfloor \qquad \text{(numeral nested)}$$

with $\lfloor x \rfloor$ as the largest integer not greater than x. That is, "numeral nested" assumes that subjects pick the first digit (of the number of tokens to donate) first and the second digit last. The other nested logit model is based on the ratio of payoffs π_i and π_j between dictator and recipient, respectively. Andreoni and Miller (2002) found that this ratio would be of significant relevance for a fair share of the subjects.

$$s \text{ and } s' \text{ are in the same nest} \quad \Leftrightarrow \quad \lfloor \pi_i(s)/\pi_j(s) \rfloor = \lfloor \pi_i(s')/\pi_j(s') \rfloor \qquad \text{(ratio nested)}$$

The formal specification of nested logit models is standard, see e.g. McFadden (1984).

Definition 4.1 (Nested logit)**.** If $(B_r)_{r\in R}$ denotes a partition of S_1 into nests, then the probability of choosing $s \in B_r \subset B$ is

$$\sigma(s) = \frac{\exp\left\{\lambda u(s|\alpha,\beta)/\rho\right\}}{\exp\{I_r\}} \cdot \frac{\exp\{\rho I_r\}}{\sum_{t\in R}\exp\{\rho I_t\}} \tag{6}$$

with inclusive values $I_r = \ln \sum_{s' \in B_r} \exp\left\{\lambda u(s'|\alpha,\beta)/\rho\right\}$ for all $r \in R$.

In relation to nested logit, the nests B_r in ordered GEV models are overlapping, but aside from this, the definition is fairly similar.

Definition 4.2 (Ordered GEV)**.** Using bandwidth $M \in \mathbb{N}_0$, $\rho \in [0,1]$, and weights $w_m \geq 0$ for all $m = 0,\ldots,M$ such that $\sum_{m=0}^{M} w_m = 1$, the choice probabilities are

$$\sigma(s) = \sum_{r=s}^{s+M} \frac{w_{r-s} \exp\left\{\lambda u(s|\alpha,\beta)/\rho\right\}}{\exp\{I_r\}} \cdot \frac{\exp\{\rho I_r\}}{\sum_{t=0}^{B+M} \exp\{\rho I_t\}} \tag{7}$$

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 $r \in \{0,\ldots,B+M\}$ and nests $B_r = \left\{s \in \{0,1,\ldots,B\} \mid r - M \leq s \leq r\right\}$.

Intuitively, player i first picks a neighborhood B_r, $r \in \{0,\ldots,B+M\}$, and second picks a strategy $s \in B_r$ in this neighborhood. Every strategy belongs to $M+1$ neighborhoods. The probability of choosing s conditional on having chosen B_r is the first factor above, and the probability of choosing nest B_r is the second factor above. These probabilities are aggregated over all neighborhoods containing s.

Small (1987, Prop. 1) shows that ordered GEV is a GEV model indeed, which implies that it is consistent with random utility maximization and a special case of the quantal response framework defined by McKelvey and Palfrey (1995). In addition, ordered GEV reduces to multinomial logit if $\rho = 1$ or $M = 0$, and Small (1987, Prop. 2) shows that the random utility components ε_s and $\varepsilon_{s'}$ are stochastically independent if $|j-k| > M$. We use the bandwidth $M = B/2$ rounded up to the nearest even (but any M that is sufficiently large would do similarly), and Gaussian weights

$$w_m = f_{\mathcal{N}(M/2,\sigma^2)}(m) \Big/ \sum_{m=0}^{M} f_{\mathcal{N}(M/2,\sigma^2)}(m), \tag{8}$$

where $f_{\mathcal{N}(\mu,\sigma^2)}$ denotes the density of the normal distribution with mean $M/2$ and variance σ^2 (the latter is estimated from the data). Finally, to improve comparability with the other models, which require four parameters per type, we assume that ρ is constant across types.

The whole set of parameter estimates can be found in the supplementary material again. Table 3 summarizes their respective measures of validity. The observations can be summarized as follows.

Table 3: Internal and external validity of the GEV models

	Number of types						
	One	Two	Three	Four	Five	Six	Seven
Numeral	3932	3877	3028	2923	2854	2833	2800
nested logit	1337	1097	1207	1236	1192	1191	938
Ratio	4294	3801	3206	3031	3106	2957	2927
nested logit	1156	1083	950	899	966	942	916
Ordered	4300	3087	3034	2857	2805	2708	2655
GEV	1195	843	891	872	858	818	807

Note: As in Table 1, the top row (per model) contains the BIC $= -LL + (\#Pars)/2 \cdot \ln(\#Obs)$ measure of internal validity and the bottom row contains the $-LL$ measure of external validity.

Result 4.3. *The best random utility model relaxing IIA is ordered GEV. Its validity overall is similar to that of random coefficient modeling, and it is better for small numbers of types. The "control model" numeral nested logit has fairly high internal validity but lacks external validity (as anticipated).*

5 Analysis of public goods contributions

To verify the robustness of the above results, we repeat the procedure in an analysis of contributions to linear public goods. The net transfers induced by contributions to public goods are comparable to donations in dictator games, and if players have linear utilities, these games are essentially equivalent in that both games induce best responses that are independent of the opponents' choices. A difference between dictator games and public goods games persists even in this case, however. In relation to dictator games, public goods games have it that small increases of altruism tend to induce comparably large increases of contributions. The reason is that in typical public goods games, there are several recipients of one's contribution and the implicit exchange rates tend to be more favorable.

An experiment analyzing the relevance of exchange rates and the consistency of decisions in this context has been reported by Goeree et al. (2002). In particular, their

experimental treatments vary the group size n as well as external returns τ_E and internal returns τ_I of individual contributions, while the costs $\tau_K = .05$ of contributions were held constant. Using N to denote the set of players and $S_i = \{0,\ldots,25\}$ as the strategy set for all $i \in N$, the payoff function of i in their experiment was

$$\pi_i(s) = \tau_K \cdot (25 - s_i) + \tau_I s_i + \tau_E \sum_{j \neq i} s_j. \tag{9}$$

Figure 2 provides an overview of the treatment parameters and the results of the experimental results. Goeree et al. (2002) estimated multinomial logit models of random utility, using linear and Cobb-Douglas functions utility functions. We will consider more general n-player CES aggregators similar to above.[5]

$$u_i = \left((1-\alpha)\pi_i^\beta + \tfrac{\alpha}{|N|-1}\textstyle\sum_{j\neq i}\pi_j^\beta\right)^{1/\beta} \tag{10}$$

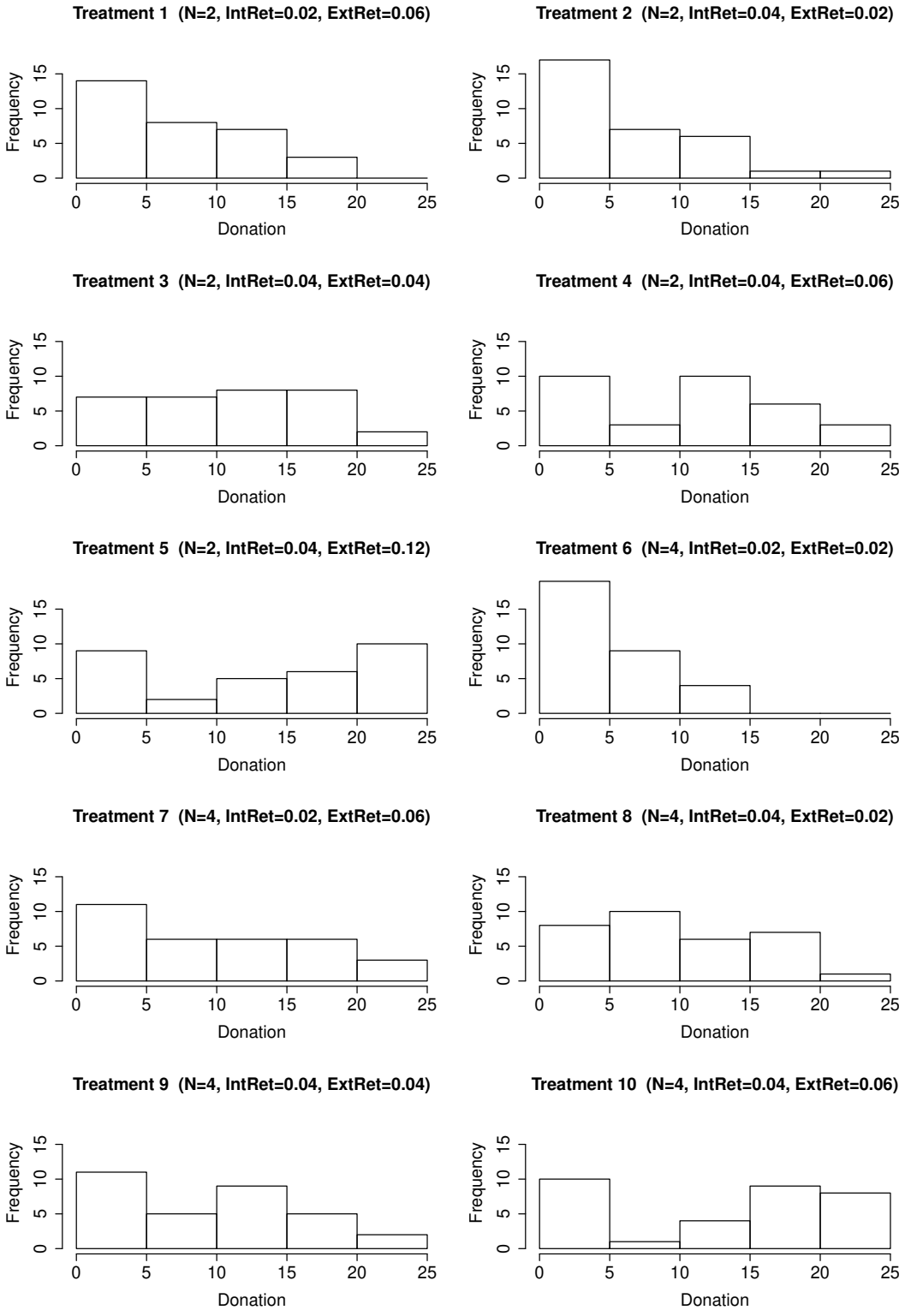
Random utility modeling (i.e. multinomial logit) was also applied by Anderson et al. (1998) to standard public goods games, by Offerman et al. (1998), Myatt and Wallace (2008), and Choi et al. (2008) to threshold public goods games, and by Willinger and Ziegelmeyer (2001) and Yi (2003) to nonlinear games. A model of random behavior was estimated by Bardsley and Moffatt (2007). Only the latter control for subject heterogeneity through finite mixture modeling. An analysis of the validity of these alternative approaches has not been reported yet, and to my knowledge, random coefficient models or GEV models have not been considered at all. Recall that these neglected models proved most valid in our analysis of dictator games.

Our analysis of public goods games mimics the above analysis of dictator games in virtually all aspects. The only notable difference is that we need to consider mutual (quantal) responses following for example McKelvey and Palfrey (1995). Using the computational simplification described by Goeree et al. (2002, Footnotes 20,21), equilibria can be computed straightforwardly.[6] This applies equally to equilibria in

[5]The CES utility is appropriate also in the case of public goods, as it contains "conditional cooperators" (Leontief preferences) and "free riders" (egoists) as special cases. These types have been identified repeatedly in the literature (Keser and van Winden, 2000; Fischbacher et al., 2001).

[6]That is, we assume that players choose best/quantal responses to the expected contributions of the opponents, and in equilibrium, they have rational expectations.

Figure 2: Treatment parameters and data of Goeree et al. (2002)



responses based on random behavior and random coefficients. As for the linear (tobit) baseline model, the independent variables are the treatment parameters again, i.e. group size, external return, and internal return.

Table 4 summarizes the measures of validity, and the supplementary material contains the whole set of parameter estimates. The main results are somewhat surprising and can be summarized as follows.

Result 5.1. *Only random utility models improve upon tobit regression in terms of internal validity, and only ordered GEV does so also in terms of external validity. The random coefficient model performs worst, as it does not even meet the lower benchmark.*

The fact that ordered GEV performs best confirms our observations from the dictator game. It scores about 50% internal as well as external validity in relation to the benchmarks, which is less than it did in the dictator game. This is a consequence of the comparably small sample size in the experiment of Goeree et al. (2002), which makes the upper benchmark particular tough to reach. Aside from this, the most interesting observation seems to be the dismal performance of the random coefficient model. It implies that the random coefficient model is not generally valid, and that its validity in dictator games may be coincidental. The underlying issue can be explained as follows. In the dictator game experiment the donation efficiencies ranged from $1/3$ to 3 (i.e. up to \$3 transfer resulted from a donation of \$1). In the public goods experiment, the donation efficiencies ranged from 1 to 18, i.e. donations were more efficient overall and much more efficient at the upper bound. When donations are that efficient, then small increases in α imply comparably large increases in the utility maximizing donation. In turn, the optimal donation is fairly sensitive with respect to α in the public goods experiment, and in this sense it poses a tougher test for constance of the altruism parameter α than the dictator game experiment.

To illustrate this, I computed the $\alpha' = \alpha/(1-\alpha)$ (see Def. 3.2) that explain the mean observation in all treatments (using the estimated β). One might expect these α' to be fairly constant. In the dictator game experiment, the ratio of the highest α' to the lowest α' in all treatments is 3.08, and in the public goods experiment it is 9.05.[7]

[7]The respective $\alpha' = \alpha/(1-\alpha)$ are $(3.141, 1.019, 2.711, 1.245, 2.802, 1.309, 1.911, 2.144)$ for the

Table 4: Internal and external validity of the models in public good games

(a) The linear behavioral model Eq. (2)

	Benchmarks		Number of types			
	Lower	Upper	One	Two	Three	Four
Internal	1043	787	1002	915	921	935
External	313	234	304	280	283	283

Note: Internal validity is BIC $= -LL + (\#Pars)/2 \cdot \ln(\#Obs)$ of the model fitted to the whole data set, external validity is $-LL$ in treatments 8-10 of the model fitted to the restricted sample (1–7).

(b) The basic structural models

	Number of types			
	One	Two	Three	Four
Random behavior	1123	988	999	1000
	350	327	323	313
Random coefficient	1179	1181	1115	1129
	480	480	393	393
Random utility	1018	936	939	915
	311	294	296	288

(c) The GEV models

	Number of types			
	One	Two	Three	Four
Numeral nested logit	1020	957	917	922
	310	293	287	285
Ratio nested logit	1006	957	916	922
	308	293	287	288
Ordered GEV	1008	932	931	908
	306	288	285	275

Thus, to explain the public goods data, highly variable α are required in random coefficient models, and in turn the actual structure cannot be described in terms of α. As we can see in Table 4, random behavior and in particular random utility models are much more suitable to capture the structure of behavior in public goods games, and thus also overall.

6 Conclusion

The paper compared the validity of econometric models to explain observations from two of the most widely researched experimental games. Our analysis utilized data sets from experiments designed to understand structure and consistency of individual decisions in dictator games (Andreoni and Miller, 2002) and public goods games (Goeree et al., 2002). The analysis covered atheoretic regression and various structural models, including the largely overlooked ordered GEV model (Small, 1987) and a "control model" relaxing IIA based on seemingly arbitrary nesting (based on numeral digits).

It was found that random utility modeling tends to be more robust than random coefficient and random behavior modeling, and of the random utility models considered, ordered GEV is most appropriate to explain the data, while the "control model" numeral nested logit indeed has low external validity as expected (in particular in dictator games). This shows that choosing the appropriate model structure has to be discussed in more detail than it is done in the current literature, where the various modeling approaches simply coexist. As for altruistic giving in dictator and public goods games, ordered GEV seems to be an appropriate model, but to my knowledge, no such results exist in alternative contexts.

In addition, our analysis confirmed the general suspicion that regression analyses may lack external validity. Regression analyses misrepresent the underlying patterns in dictator games and therefore postulate results that do not continue to hold in related

dictator game treatments and $(0.475, 0.49, 0.24, 0.161, 0.078, 0.483, 0.163, 0.164, 0.082, 0.054)$ for the public goods game treatments. These are the lower bounds for α' to explain the mean contribution rounded to the nearest integer.

dictator games. In this sense we can say that regression analysis does not let the data speak for itself, but that it squeezes the data into a linear (or non-linear) form that would fit only coincidentally. To be sure, structural models do not lead to easily digestible conclusions, as their estimates are utility functions and noise parameters, but that it is exactly the key: easily digestible linear effects are not robust and structurally wrong. In turn, structurally wrong structural models are not robust either, but as our results show for dictator and public goods game, such models do usually not have internal validity to begin with, and random utility models fit robustly in both cases.

References

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

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.

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.

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

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

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

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

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. (2008). Mixture models of choice under risk. *Journal of Econometrics (forthcoming)*.

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

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

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., 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.

Haile, P., Hortacsu, A., and Kosenok, G. (2008). On the empirical content of quantal response equilibrium. *American Economic Review*, 98(1):180–200.

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

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. P. (2010). 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.

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

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.

McFadden, D. (1984). Econometric analysis of qualitative response models. *Handbook of econometrics*, 2:1395–1457.

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

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

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.

Peel, D. and MacLahlan, G. (2000). *Finite Mixture Models*. Wiley interscience.

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.

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.

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

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.

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.